특집논문 (Special Paper)
방송공학회논문지 제26권 제2호, 2021년 3월 (JBE Vol. 26, No. 2, March 2021)
https://doi.org/10.5909/JBE.2021.26.2.125
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

저계수 행렬 근사 및 CP 분해 기법을 이용한 CNN 압축

문 현 철[a)], 문 기 화[a)], 김 재 곤[a)‡]

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

HyeonCheol Moon[a)], Gihwa Moon[a)], and Jae-Gon Kim[a)‡]

요 약

최근 CNN(Convolutional Neural Network)은 영상 분류, 객체 인식, 화질 개선 등 다양한 비전 분야에서 우수한 성능을 보여주고 있다. 그러나 많은 메모리와 계산량이 요구되어 모바일 또는 IoT(Internet of Things) 장치와 같은 저전력 디바이스에 적용하기에는 제한이 따른다. 이에, CNN 모델의 임무 성능을 유지하면서 네트워크 모델을 압축하는 연구가 진행되고 있다. 본 논문에서는 행렬 분해 기술인 저계수 행렬 근사(Low-rank approximation)와 CP(Canonical Polyadic) 분해 기법을 결합한 CNN 모델 압축 기법을 제안한다. 제안기법은 하나의 행렬 분해 기법만을 적용하는 기존의 기법과 달리 CNN의 계층 유형에 따라 두 가지 분해 기법을 선택적으로 적용하여 압축 성능을 높인다. 제안기법의 성능 검증을 위하여 영상 분류 CNN 모델인 VGG-16, ResNet50, 그리고 MobileNetV2 모델을 압축하였고, 계층 유형에 따라 두 가지의 분해 기법을 선택적으로 적용함으로써 저계수 행렬 근사 기법만 적용한 경우 보다 1.5 ~ 12.1 배의 동일한 압축률에서 분류 성능이 향상됨을 확인하였다.

Abstract

In recent years, Convolutional Neural Networks (CNNs) have achieved outstanding performance in the fields of computer vision such as image classification, object detection, visual quality enhancement, etc. However, as huge amount of computation and memory are required in CNN models, there is a limitation in the application of CNN to low-power environments such as mobile or IoT devices. Therefore, the need for neural network compression to reduce the model size while keeping the task performance as much as possible has been emerging. In this paper, we propose a method to compress CNN models by combining matrix decomposition methods of LR (Low-Rank) approximation and CP (Canonical Polyadic) decomposition. Unlike conventional methods that apply one matrix decomposition method to CNN models, we selectively apply two decomposition methods depending on the layer types of CNN to enhance the compression performance. To evaluate the performance of the proposed method, we use the models for image classification such as VGG-16, RestNet50 and MobileNetV2 models. The experimental results show that the proposed method gives improved classification performance at the same range of 1.5 to 12.1 times compression ratio than the existing method that applies only the LR approximation.

Keyword : CNN, Neural Network Compression, Low-Rank Approximation, Canonical polyadic decomposition

Ⅰ. 서 론

최근 CNN(Convolutional Neural Network)을 기반으로 하는 인공신경망은 영상인식 및 화질개선 등 다양한 컴퓨터 비전 응용에서 뛰어난 성능을 보이고 있다. CNN은 ILSVRC(ImageNet Large Scale Visual Recognition Challenge)에서 우수한 분류 성능을 달성하였으며 현재 이미지 인식 분야에서 대표적인 신경망으로 사용되고 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 증가하였다. 이에 연산속도나 메모리가 제한된 모바일 및 IoT 기기에 CNN 기반의 인공신경망 적용하기에는 제약이 따른다. 예를 들어, 이미지 분류 모델인 VGG-16은 500MB가 넘는 파라미터들을 가지고 있어서 추론할 때마다 이를 접근하고 저장해야 하는데 이는 저전력 기기에 큰 부담을 준다. 이러한 문제들을 해결하기 위해 기존의 학습된 네트워크 모델의 성능을 최대한 유지하면서 모델의 크기를 줄이는 인공신경망 경량화 연구가 중요해지고 있다[1]. 이와 관련하여 인공신경망 압축을 위한 MPEG NNR(Compression of Neural Network for Multimedia Content Description and Analysis) 표준화가 진행되고 있으며 마무리 단계에 있다[2].

인공신경망 모델의 크기 및 연산량을 줄이는 연구는 크게 연산에 효율적인 인공신경망 구조를 설계하는 인공신경망 경량화 기법과 학습된 가중치 등 모델 파라미터를 압축 표현하는 인공신경망 압축 기법으로 나눌 수 있다. 인공신경망 경량화 기법은 파라미터 축소 및 모델 성능을 개선할 수 있는 신경망 구조를 개선하는 연구와 각 계층의 구조에서 연산량을 효율적으로 줄이는 연구 등이 있다[3][4]. 인공신경망 압축 기법은 인공신경망의 구조적 설계가 아닌 이미 학습된 인공신경망 파라미터의 표현 및 크기를 줄이는데 중점을 두고 있다. 대표적인 인공신경망 압축 기법으로는 매우 작은 값을 가지는 가중치의 경우 모델의 성능에 미치는 영향이 적다고 판단하고 연결을 끊는(즉, 해당 가중치 값을 0으로 변경) 가지치기(pruning) 기법과 일반적으로 부동소수점으로 표현되는 가중치를 특정 비트 수로 표현하여 실제 모델의 저장 크기를 줄이는 양자화 기법이 있다[5][6]. 또한 가중치를 0과 1로 표현하여 모델 임무 성능의 손실을 어느 정도 범위내로 유지해 주면서 모델의 저장 크기를 크게 줄여주는 이진화 기법과 각 계층의 2차원 이상의 가중치 행렬을 2개 이상의 행렬로 분해하여 파라미터 수를 줄이는 행렬분해 기법 등이 있다[7]. 이러한 압축 기법들을 CNN 모델 등에 적용함으로써 기존의 모델 대비 모델의 크기를 줄일 수 있을 뿐만 아니라 실제 추론 과정에서의 연산량 또한 줄일 수 있다.

본 논문에서는 저계수 행렬 근사(LR: Low-Rank approximation)와 CP(Canonical Polyadic) 분해 기법을 결합한 행렬 분해 기법을 제안한다[8-11]. 즉, CNN 모델의 계층의 유형에 따라 LR과 CP를 선택적으로 적용함으로써 단일 행렬 분해 기법을 적용하는 기존 기법 대비 압축 성능을 개선하고자 한다. 제안기법의 성능 검증을 위해서 이미지 분류 CNN 모델인 VGG-16, ResNet50 그리고 MobileNetV2에 적용하여 동일한 모델 압축율 범위에서 기존 모델 대비 모델의 Top-5 성능을 측정한다.

본 논문의 구성은 다음과 같다. 2장에서는 행렬 분해 기법인 저계수 행렬 근사 기법과 CP 분해 기법에 대해 설명한다. 3장에서는 실험 적용을 위한 제안기법을 설명하고 4장에서는 상세 실험내용과 각 모델에서의 실험결과에 대해 분석하며, 마지막으로 5장에서 결론을 맺는다.

Ⅱ. 행렬 분해 기법

행렬분해 기법은 2차원 이상의 행렬로 표현되는 각 계층의 가중치 행렬을 2개 이상의 행렬로 분해함으로써 가중치 파라미터 수 및 연산량을 줄이는 기법이다. 대표적인 행렬

a) 한국항공대학교 항공전자정보공학부(Korea Aerospace University, School of Electronics and Information Engineering)
‡ Corresponding Author : 김재곤(Jae-Gon Kim)
E-mail: jgkim@kau.ac.kr
Tel: +82-2-300-0414
ORCID: https://orcid.org/0000-0003-3686-4786
※ 이 논문의 연구결과 중 일부는 한국방송·미디어공학회 “2020년 추계학술대회”에서 발표한 바 있음.
※ This work has supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1F1A1068106).
· Manuscript received January 8, 2021; Revised March 3, 2021, Accepted March 3, 2021.

분해 기법으로는 2차원 행렬을 SVD(Singular Value Decomposition) 분해 기법을 이용하여 2개의 행렬로 분해하는 저계수 행렬 근사 기법[8]과 3차원 이상의 행렬을 다수의 랭크(rank)-1 텐서(tensor)의 선형결합 형태로 분해하는 CP 분해 기법[10] 등이 있다.

저계수 행렬 근사 기법은 2차원 행렬을 SVD 분해 기반으로 2개의 2차원 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_i는 각 CNN 모델의 i번째 계층의 가중치 행렬을 의미하며, U_i, V_i^T는 i번째 층의 가중치 행렬 W_i로부터 분해되는 2개의 행렬을 의미한다. 또한 R은 분해의 계수인 Rank값을 나타낸다.

CP는 3차원 이상의 고차원 텐서들을 N 개의 Rank-1 텐서들의 선형결합으로 분해한다. 예를 들어, 식 (2)와 같이 3차원 텐서 X를 R 개로 이루어진 3개의 Rank-1의 텐서로 분해할 수 있다. 여기서 R은 해당 기법을 적용하기 위해 입력되는 하이퍼파라미터(hyper-parameter)이며, R값에 따라 압축률이 결정된다.

$$\begin{gathered} \mathrm{X} \equiv \sum_{r=1}^{R} a_r \otimes b_r \otimes c_r \\ \otimes : outer\ \ product \\ a \otimes b = ab^T : where\ a,b\ are\ rank-1\ tensors \end{gathered} \tag{2}$$

본 논문에서는 4차원 기반의 컨볼루션(convolution) 가중치 텐서들을 CP 분해 기법을 적용하여 압축한다[5]. 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 3개의 계층으로 분해하는 기법을 제시한다.

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

그림 1은 식 (3)을 그림으로 나타낸 것이다. 그림 1에서 의 H와 W는 현재 계층의 입력으로 들어갈 특징 맵의 가로 및 세로 크기이며, 이 때 출력되는 특징 맵은 Stride 계산에 따라 가로 세로의 크기가 변할 수 있으므로 가로 및 세로를 H', W'로 각각 표현하였다. 그림1-(a)는 분해되기 전 기존의 $D\times D$ 크기의 필터를 가지고 있는 컨볼루션 층에 대한 예시를 나타낸다. 그림1-(b)는 1-(a)의 컨볼루션 층에 CP분해된 계층의 예시를 나타내며, 3개의 화살표는 각각 순서대로 식 (3)의 $U_{s,r}^{(1)}$, $U_{r,j,i}^{(2)}$, $U_{r,t}^{(3)}$를 의미한다. 여기서 $U_{s,r}^{(1)}$, $U_{r,j,i}^{(2)}$, $U_{r,t}^{(3)}$는 각각 $S\times R$, $R\times D\times D$, $R\times T$ 크기를 가진다. 이 때, S와 T는 각각 입력과 출력의 채널 수, $D\times D$는 2D 컨볼루션 필터 크기를 의미한다. 따라서, $U_{s,r}^{(1)}$, $U_{r,t}^{(3)}$은 필터 크기가 1×1인 Point-wise(PW) 컨볼루션 층의 형태로 구성되며, $U_{r,j,i}^{(2)}$는 필터 크기가 $D\times D$인 Depth-wise (DW) 컨볼루션 층으로 구성된다. 따라서, CP분해는 하나의 $D\times D$ 컨볼루션 층을 2개의 Pointwise, 1개의 Depth-wise 컨볼루션 층으로 분해한다.

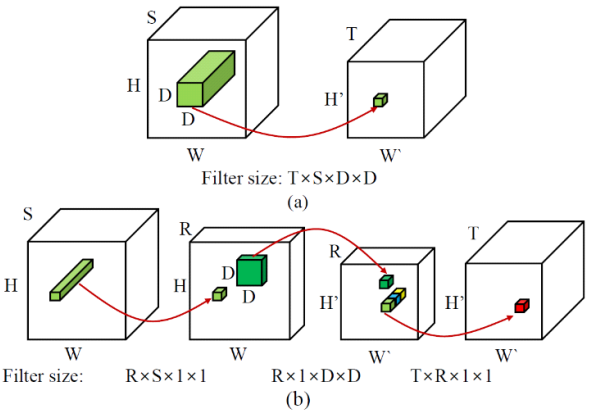

그림 1. CP 분해 기법 (a) 컨볼루션 층 (b) CP 분해된 컨볼루션[9]
Fig 1. CP decomposition (a) Convolutional layer (b) CP decomposed convolutional layer

Ⅲ. 제안기법

CNN 에는 여러 종류의 컨볼루션 타입이 존재하고 각 모델마다 서로 다른 계층별 컨볼루션 파라미터 분포를 가지고 있다. 컨벌루션 계층의 종류에는 흔히 사용되는 2D 컨볼루션, DW 컨볼루션, 그리고 PW 컨볼루션 등이 있다. DW

컨볼루션은 각 채널마다의 공간적인 특징를 추출하기 위한 것이며 각 채널마다 하나의 필터만 존재하게 된다. 반면에 PW 컨볼루션은 1x1 크기로 고정된 필터를 사용한다. DW와는 달리 공간적인 특징를 다루지 않고 채널들에 대해서만 스케일링(scaling) 연산을 수행한다.

본 논문에서는 컨볼루션 층을 저계수 행렬 근사(LR)) 및 CP 분해 기법을 결합한 행렬 분해 기법을 제안한다[9-10]. 여기서, 컨볼루션 층의 가중치 행렬인 경우 4차원 형태 ($S\times T\times D\times D$)로 구성되어 있기 때문에 2차원 행렬을 2개로 분해하는 저계수 행렬 근사 기법을 적용함에 제한이 따른다. 따라서, 저계수 행렬 근사 기법을 적용하기에 앞서 2차원으로 재배열 ($SD\times DT$) 한 뒤 분해 기법을 적용하며, 이는 PW 컨볼루션 층(D=1)에도 동일하게 적용된다. 표 1은 제안된 기법의 각 실험 별 조건을 나타낸 것이다. 기존의 행렬 분해로 CNN을 압축하는 기법들은 계층의 유형에 상관없이 단일 분해 기법만을 적용한다. 제안기법은 성능 향상을 위해 각 실험에서는 계층의 유형에 따라 적용하는 행렬 분해 기법들을 다르게 적용한다. 완전연결(FC: Fully-Connected) 층의 경우 모든 실험 조건에서 저계수 행렬 근사(LR) 기법을 적용하였고, 컨볼루션 층에는 2개의 기법을 모두 선택적으로 적용한다 (Test 1&2). 특히 Test 3에서는 각 계층에서 계층의 유형에 따라 행렬 분해 기법을 적용하지 않는 반면에 식 (4)로 주어진 복원 손실(reconstruction loss)에 따라 행렬 분해 기법을 선택적으로 적용한다.

$$L_i = \| W_i - W_i^{'} \|^2 \quad (4)$$

여기서 W_i는 각 CNN 모델의 i번째 계층의 가중치 행렬을 의미하며, $W_i^{'}$는 W_i로부터 분해된 행렬을 다시 복원한 행렬을 의미한다. 즉, 두 개 행렬의 차원은 동일하다. 따라서, L_i는 원래의 가중치 행렬과 행렬분해 후 복원한 행렬의 평균 제곱 오차 값으로 계산된다. 즉, Test 3는 특정 계층에 낮은 순위 행렬 근사 방법과 CP 분해 방법을 각각 적용하여 오차가 적은 행렬분해 기법을 적용하는 것이다.

표 2는 실험에 사용된 모델들의 계층별 파라미터 분포를 나타낸다. VGG-16과 MoblieNetV2의 경우 ResNet50 과 달리 2D 컨볼루션 파라미터와 PW 및 DW 컨볼루션 파라미터의 분포의 차이가 크다. 이러한 모델 들은 계층의 유형에 따른 서로 다른 행렬 분해방법을 적용하는 것은 큰 효과가 없으므로 계층별 파라미터의 분포가 유사한 Resnet50 모델에서만 계층의 유형에 따른 각각의 분해 기법을 적용하여 실험을 진행하였다(Test 1~3).

표 1. 실험조건
Table 1. Experimental conditions

	Layer Type		
	2D Conv.	PW Conv.	FC
LR Baseline	LR	LR	LR
Test 1	CP	CP	
Test 2	CP	LR	
Test 3	Hybrid (CP & LR)		

표 2. 각 모델의 계층별 파라미터 분포
Table 2. Parameter distributions of layer types in each model

Model	General convolution(2D)	PW & DW convolution	FC layer
VGG-16	11%	0%	89%
ResNet50	44%	47%	9%
MobileNetV2	1%	63%	36%

Ⅳ. 실험결과

본 논문의 실험에서는 영상 분류 CNN 모델인 VGG-16, ResNet50, MobielNetV2에 대하여 원본 모델 및 제안한 인공신경망 압축 기법이 적용된 모델들의 성능을 비교하였다.

표 3은 실험에 사용한 CNN 원본 모델의 영상 분류 성능이다[7]. 표 1에서의 Top-5 Accuracy는 모델로부터 예측된 클래스 중 상위 5개에 정답이 있을 경우의 정확도를 의미한다. 본 논문에서의 분류 성능을 측정하기 위해 사용된 데이터 셋은 ILSVRC 2012의 검증(validation) 데이터 셋의 50,000개의 영상이며[12], 입력되는 영상의 크기는 224×224이다. 실험에서는 단순히 행렬 분해 기법만을 적용하면 분류 성능의 손실이 발생하므로 별도의 재학습(Re-training)을 적용하였으며, 재학습을 위해 ILSVRC 2012의 학습 데이터 셋 500,000개의 영상을 사용하였다.

표 3. 실험에 사용된 CNN 원본 모델의 성능(Anchor accuracy)
Table 3. Performances of the CNN original models used in the experiment (Anchor accuracy)

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

표 4는 계층 타입에 따른 압축율을 나타내며, 모든 실험에서 동일하게 적용하였다. 본 논문의 실험을 위해 첫 계층을 제외한 모든 컨볼루션 층에 적용하였으며, 같은 계층의 유형에서는 동일한 압축율을 적용하였다. 예를 들어, VGG-16의 2D 컨볼루션 13계층을 4.4배 압축하기 위하여 첫번째 컨볼루션 층을 제외한 나머지 12개 계층은 4.4배 압축할 수 있는 행렬 분해 계수 Rank 값을 각 계층별로 구하여 압축한다.

표 4. 계층 유형에 따른 압축율 (R값)
Table 4. Compression ratio (R value) according to layer type

Model	General convolution (compression ratio)	PW & DW convolution (compression ratio)	FC layer (compression ratio)
VGG-16	44% (4.4x)	-	9% (10.8x)
ResNet50	38% (2.6x)		100% (1x)
MobileNetV2	-	60% (1.6x)	80 (1.2x)

표 5는 행렬 분해 기법을 적용한 경우의 압축 성능 실험 결과이다. 각 압축 기법을 적용한 압축율은 압축된 모델의 Top-5 분류 성능이 원본 모델 대비 3% 미만의 손실일 때를 기준으로 분류 성능을 비교하였다. 표 5에서와 같이 행렬 분해 기법을 적용한 실험에서 분류의 임무 성능 손실 없이 주어진 모델을 1.5 ~ 12.1배 압축함을 확인하였으며, VGG-16에서는 기존의 가지치기 기법[7] 대비 약 1.3배 더 높은 압축율을 얻었다. 더불어, 계산량도 역시 기존 모델 대비 약1.2 ~ 2.4배 향상됨을 확인하였다.

반면에, 모델 별 동일한 압축율에서 성능을 비교한 결과 VGG-16에서는 Test 1, 그리고 ResNet50에서는 Test 2~3에서 모든 계층에 저계수 행렬 근사 기법만을 적용한 기법보다 계층의 유형에 따라 LR과 CP기법을 선택적으로 적용한 제안기법에서 성능 향상이 있음을 확인하였다.

표 5. 계층별 압축 기법 적용에 따른 분류 성능 실험결과
Table 5. Experimental results on the classification performance of the applied methods depending on layer types

Model	Test	Top-1/5 Accuracy (%)	Org. Top-1/5 Accuracy (%)	Compression ratio(x)	
				Memory	Comp
VGG-16	LR	67.35/87.92	70.62/90.05	12.1	2.4
	Test 1	67.89/88.11			
	Test 3	68.01/88.20			
	Hans[7]	67.83/88.05		9.0	-
ResNet50	LR	70.94/89.44	74.97/91.93	2.9	2.1
	Test 1	70.12/88.79			
	Test 2	70.79/89.87			
	Test 3	70.98/89,90			
MobileNetV2	LR	67.83/88.20	71.49/90.06	1.5	1.2
	Test 1	67.25/88.07			

표 6은 Test 3에서의 각 기법의 비율을 나타낸 것이다. 모델 실험에서 2D 컨볼루션 계층에서는 CP 분해 기법을 적용할 때 대비 LR 근사 기법 보다 높은 성능을 나타냈다. 또한, 계층의 유형에 상관없이 손실 값으로 분해 기법을 선택적으로 적용한 Test 3가 VGG-16과 ResNet50 실험 중 성능이 제일 좋음을 확인하였다. 반면에 MobileNetV2에서는 LR 근사 기법만을 적용한 기존의 방법보다 서로 다른 분해 기법을 사용한 Test 3에서 낮은 성능을 보여주며, 이는 PW 및 DW 컨볼루션 층에는 LR 근사 기법이 더 성능이 좋음을 보여준다. 실제로 표 6에서 2D 컨볼루션층에는 CP 기법이 LR 기법 대비 약 2~3배 넘게 선택되었고, PW 및 DW 컨볼

표 6. Test 3(하이브리드 방법)에서의 각 기법 (LR/CP) 비율
Table 6. Percentage of each method in Test 3 (hybrid approach)

Model	General convolution (2D)		PW & DW convolution	
	LR	CP	LR	CP
VGG-16	25%	75%	-	
ResNet50	31%	69%	82%	18%
MobileNetV2	-		89%	11%

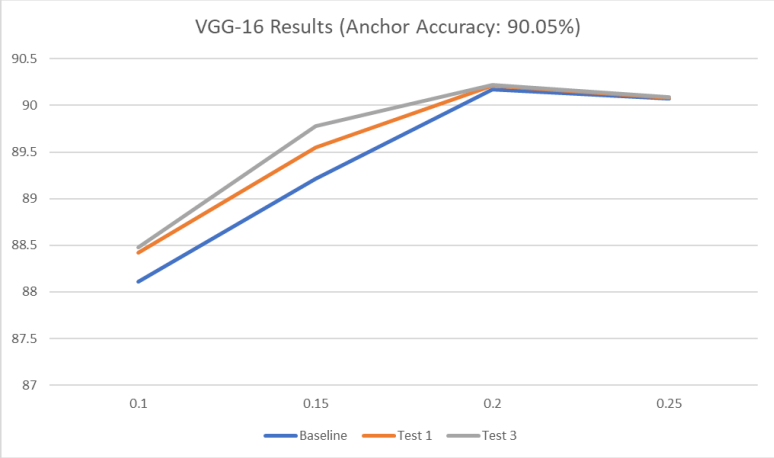

그림 2. VGG-16 에서의 압축률에 따른 분류 성능
Fig. 2. Classification performance of VGG-16 in the range of compression

루션층에서는 LR 기법이 CP 기법 대비 약 5~8배 더 넘게 선택됨을 확인하였다.

그림 2는 VGG-16에서의 각 실험 별 성능 압축률에 따른 성능을 나타낸 것이다. 컨볼루션 층에 CP 분해 기법을 적용한 Test 1이 LR 근사 기법만을 사용한 것보다 압축율 전체 구간에서 분류 성능이 더 좋음을 확인하였으며, 손실 값에 의해 분해 기법을 선택적으로 적용한 Test 3가 Test 1 보다 성능의 향상이 있음을 확인하였다.

V. 결 론

본 논문에서는 행렬 분해 기법인 저계수 행렬 근사(Low-Rank approximation)와 CP(Canonical Polyadic) 분해 기법을 선택적으로 적용하여 학습된 CNN 모델을 압축하는 기법을 제안하였다. 영상 분류의 대표적 CNN 모델인 VGG-16, ResNet50, 그리고 MobileNetV2 모델에 대하여 2 가지의 행렬 분해 기술을 컨볼루션 계층 유형에 따라 선택적으로 적용하는 제안기법의 압축 성능을 검증하였다. 실험결과 제안한 행렬 분해 기법은 단일 행렬 분해 기법을 적용하는 기존 기법 대비 3%의 미미한 분류 성능 손실 범위에서 모델 파라미터 수를 1.3 ~ 12.1 배까지 감소함으로써 압축 성능을 향상시킬 수 있음을 확인하였다.

참 고 문 헌 (References)

[1] S. Jung, C. Son, S. Lee, J. Han, Y. Kwak, and S. Hwang, “Learning to Quantize Deep Networks by Optimizing Quantization Intervals with Task Loss,” In Proc. Computer Vision and Pattern Recognition (CVPR), 2019.

[2] W. Bailer, et al, “Text of ISO/IEC DIS 15938-17 Compression of Neural Networks for Multimedia Content Description and Analysis,” ISO/IEC/JTC1/SC29/WG04, N0016, Oct. 2020.

[3] X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An Exteremey Efficient Convolutional Neural Network for Mobile Devices,” In Proc. Computer Vision and Patter Recognition (CVPR), 2018.

[4] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” arXiv preprint arXiv:1704.04861, 2017

[5] 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.

[6] H. Moon, H. Lee, and J. Kim, “Acceleration of CNN Model Using Neural Network Compression and its Performance Evaluation on Embedded Boards,” In Proc. KIBME Annual Fall Conf., Nov. 2019.

[7] 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.

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

[9] 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.

[10] H. Moon, G. Moon, and J. Kim, “Compression of CNN Using Low-Rank Approximation and CP Decomposition Methods,” In Proc. KIBME Annual Fall Conf., Nov. 2020.
[11] H. Moon, J. Kim, S. Kim, S. Jang, and B. Choi, “KAU/KETI Response to the CE-1 on Neural Network Compression: CP Decomposition of Convolutional Layers (Method 5),” ISO/IEC JTC1/SC29/WG04, m55053, Oct. 2020.
[12] Large Scale Visual Recognition Challenge 2012 (ILSVRC 2012), [Available at Online] http://www.image-net.org/challenges/LSVRC/2012/

저 자 소 개

문 현 철

- 2018년 2월 : 한국항공대학교 항공전자정보공학부 학사
- 2020년 8월 : 한국항공대학교 항공전자정보공학과 석사
- ORCID : http://orcid.org/0000-0002-1672-2345
- 주관심분야 : 비디오 부호화, 영상처리, 딥러닝

문 기 화

- 2021년 2월 : 한국항공대학교 소프트웨어학과 학사
- 2021년 3월 ~ 현재 : 한국항공대학교 항공전자정보공학과 석사과정
- ORCID : https://orcid.org/0000-0001-6727-7790
- 주관심분야 : 비디오 부호화, 영상처리, 딥러닝

김 재 곤

- 1990년 2월 : 경북대학교 전자공학과 학사
- 1992년 2월 : KAIST 전기 및 전자공학과 석사
- 2005년 2월 : KAIST 전기 및 전자공학과 박사
- 1992년 3월 ~ 2007년 2월 : 한국전자통신연구원(ETRI) 선임연구원/팀장
- 2001년 9월 ~ 2002년 7월 : Columbia University 연구원
- 2015년 12월 ~ 2016년 1월 : UC San Diego, Visiting Scholar
- 2007년 9월 ~ 현재 : 한국항공대학교 항공전자정보공학부 교수
- ORCID : http://orcid.org/0000-0003-3686-4786
- 주관심분야 : 비디오 부호화 표준, 비디오 신호처리, Immersive Video, Deep Learning