MPEG-NNR 의 지역 비선형 양자화를 이용한 CNN 압축

이정연, 문현철, 김수정, 김재곤
한국항공대학교
{leejo1998, hcmoon, sue05020}@kau.kr, jgkim@kau.ac.kr

Compression of CNN Using Local Nonlinear Quantization in MPEG-NNR

Jeong-Yeon Lee, Hyeon-Cheol Moon, Sue-Jeong Kim, and Jae-Gon Kim
Korea Aerospace University

요 약

최근 MPEG 에서는 인공신경망 모델을 다양한 딥러닝 프레임워크에서 상호운용 가능한 포맷으로 압축 표현할 수 있는 NNR(Compression of Neural Network for Multimedia Content Description and Analysis) 표준화를 진행하고 있다. 본 논문에서는 MPEG-NNR 에서 CNN 모델을 압축하기 위한 지역 비선형 양자화(Local Non-linear Quantization: LNQ) 기법을 제시한다. 제안하는 LNQ 는 균일 양자화된 CNN 모델의 각 계층의 가중치 행렬 블록 단위로 추가적인 비선형 양자화를 적용한다. 또한, 제안된 LNQ 는 가지치기(pruning)된 모델의 경우 블록내의 영(zero) 값의 가중치들은 그대로 전송하고 영이 아닌 가중치만을 이진 군집화를 적용한다. 제안 기법은 음성 분류를 위한 CNN 모델(DCASE Task)의 압축 실험에서 기존 균일 양자화를 대비 동일한 분류 성능에서 약 1.78 배 압축 성능 향상이 있음을 확인하였다.

1. 서론

CNN(Convolutional Neural Network) 기반 인공신경망은 영상 분류, 객체 인식 등 다양한 분야에서 뛰어난 성능을 보이고 있는 반면, 인공신경망 모델의 복잡도가 증가하면서 저전력 기기에서 추론을 하기에는 많은 제한이 따른다. 예를 들면 학습된 모델을 클라우드 서버에서 송/수신하거나 기기에서 모델을 저장하고 있어야 하는데, 저전력 기기에서는 계산속도와 메모리의 제한이 따른다. 따라서 최근 MPEG 에서는 NNR(Compression of Neural Networks for Multimedia Content Description and Analysis)이라는 상호운용 가능한 형태로 모델 파라미터를 압축 표현하는 표준화를 진행하고 있다[1]. NNR 에서는 인공신경망 모델 압축을 위해서 파라미터의 수를 줄이는 가지치기(pruning) 및 행렬 분해 기법과 가중치 값을 근사화 하는 양자화, 그리고 엔트로피 부호화 방법을 채택하고 있다.

본 논문에서는 MPEG-NNR 에서 CNN 모델을 압축하기 위한 기존의 지역 이진 군집화(Local Binary Clustering: LBC) 방법[2]에서 가지치기 된 모델이 입력인 경우에 대한 개선 방법인 지역 비선형 양자화(Local Non-linear Quantization: LNQ) 기법을 제시한다.

2. 지역 이진 군집화(LBC) 기법

지역 이진 군집화 기법 기법은 각 컨볼루션 층의 가중치 행렬을 일정 크기의 블록 단위로 블록 내의 가중치를 이진화 하여 압축한다[2]. 예를 들어, 그림 1 에서와 같이 균일 양자화된 가중치 값들은 K-means 양자화(이 때, k=2)를 통해 이진화된 계수 값(000111100)과 이진 코드북(Binary codebook)만으로 전송할 수 있다.

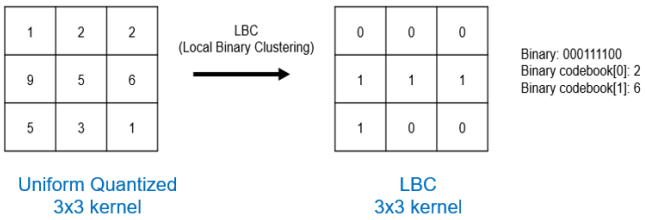

그림 1. 3x3 커널에서의 LBC 적용 예

그림 2 는 기존 LBC 기법의 부호화 과정을 나타낸 것이다. 먼저, 주어진 모델의 각 계층 내 균일 양자화가 적용된 가중치 행렬이 입력되면 해당 계층의 유형이 컨볼루션 층(Conv layer) 혹은 완전연결층(Fully Connected layer: FC layer)인지를 확인한다. 이 때, 각 계층 별로 LBC 의 적용 여부를 알려주는 지시자 LBC_Layer_flag 를 전송한다. 가중치 행렬을 LU(LBC Unit) 단위로 가중치 행렬을 분할한다. 여기서, LU 의 크기는 각 계층의 유형에 따라 다르게 정의할 수 있으며, 본 논문에서는 컨볼루션 층에서는 필터 크기로, FC 층에서는 4x4 크기로 설정하였다. 각 LU 별로 LBC 를 수행하게 되는데, 기존 균일 양자화와 LBC 의 RD(Rate-Distortion) 비용이 적은 방법을 적용한다. LBC 가 적용되면 지시자 LBC_Flag 와 코드북, 그리고 이진화된 계수 값을 전송한다.

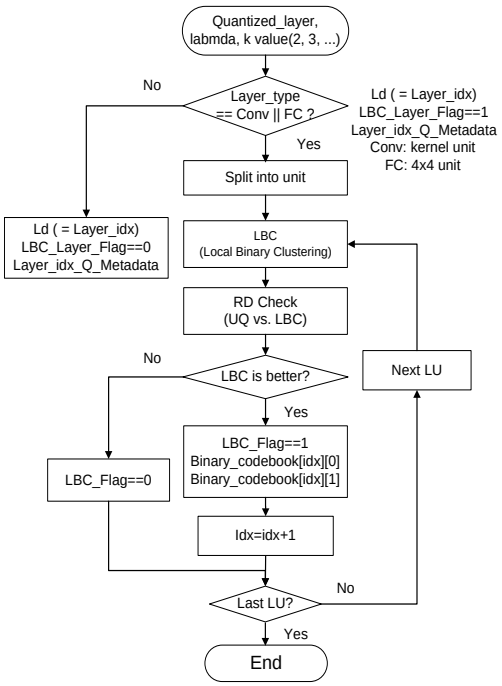

그림 2. LBC 기법 부호화 과정

3. 제안 지역 비선형 양자화(LNQ) 기법

기존의 LBC 에서는 LU 내의 양자화된 계수 값들을 이진 군집화 하였다. 그러나, 가지치기된 모델의 경우 0 을 많이 포함하고 있기 때문에 이진 군집화가 비효율적일 수 있다. 예를 들어 그림 3 은 양자화된 모델의 가중치 행렬이 입력되는 경우로 LU 내에 영 값, 양수, 그리고 음수로 이루어져 있어 이진 군집화로 적용하기에는 적절하지 않다. 따라서, 본 논문에서는 그림 3(b)처럼 0 계수들은 그대로 전송하고, 0 이 아닌 가중치들을 이진 군집화하는 지역 비선형 양자화 방법을 제안한다. 제안하는 기법은 기존의 LBC 와 동일하게 각 LU 별로 2 개의 코드북 k 을 전송하면서 양자화 오차 줄일 수 있다. 예를 들어, 그림 3(a)의 경우 SAD(Sum of Absolute Difference) 값이 7, (b)의 경우 4 이다. 따라서, 제안하는 LNQ 방법은 가지치기 모델인 경우 기존 LBC 방법보다 전송되는 정보량은 비슷하면서, 양자화 오차를 줄일 수 있다.

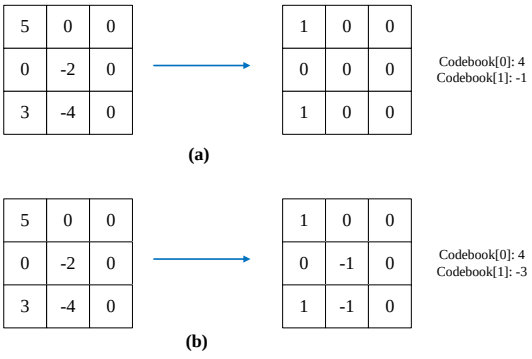

그림 3. (a) 기존 LBC 예, (b) 제안하는 LNQ 예

4. 실험결과

본 논문의 실험에 사용한 모델은 MPEG-NNR 의 유스케이스(Use Case) 중 하나인 음성 분류 DCASE 모델을 입력으로 사용하였다[3]. 제안 LNQ 의 성능을 기존의 LBC 와 균일 양자화 기법과 비교하였다.

그림 3 은 DCASE 에서의 희소도(sparsity) 85%로 가지치기된 모델의 압축율에 따른 Top-1 정확도의 실험결과이다. 이 때 희소도는 전체 가중치 수 대비 0 의 값을 가진 가중치의 비율을 의미한다. 그래프에서의 x 축은 원본 모델 대비 압축된 모델의 압축율이며, y 축은 음성 분류의 Top-1 정확도이다. 그림 3 에서와 같이 압축이 많이 된 경우 0 의 값을 제거하지 않은 기존의 LBC 보다 제안된 LNQ 가 더 뛰어난 성능을 보여줌을 확인하였다. 또한, 원본 모델 대비 성능 손실 2% 이내 조건에서 기존의 균일 양자화 대비 압축 성능이 약 1.78 배 향상됨을 알 수 있다(0.03 vs. 0.05).

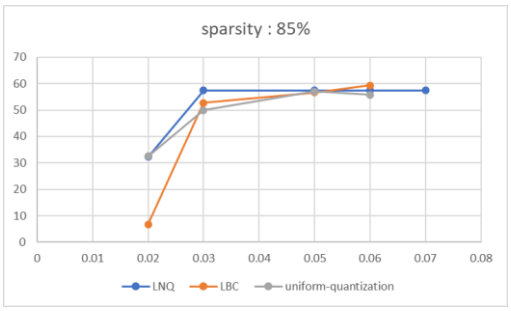

그림 3. DCASE 에서의 실험 결과 (sparsity 85%)

5. 결론

본 논문에서는 가지치기된 모델 압축에 보다 효율적인 지역 비선형 양자화(LNQ)기법을 제안하였다. 제안 기법은 0 값의 가중치는 그대로 전송하고, 0 이 아닌 가중치 만을 이진 군집화 방법으로 압축함으로써 기존의 지역 이진 군집화(LBC)의 성능을 개선하였다. 실험결과 희소성 85%로 가지치기된 음성 분류 CNN 모델인 DCASE 에서 원본 모델 성능의 2% 손실 범위내에서 기존의 균일 양자와 대비 1.78 배 압축율이 향상됨을 확인하였다. 향상이 나타남을 확인하였다.

감사의 글

이 논문은 2019 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2019-0-01351)

참 고 문 헌

[1] B. Wailer, “Working Draft 4 of Compression of neural network for multimedia content description and analysis,” ISO/IEC JTC1/SC29/WG11, N19225, Jan. 2020.

[2] H. Moon, J. Kim, S. Kim, S. Jang, and B. Choi, “KAU/KETI Response to the CE-2 on Neural Network Compression: Local Binary of Quantized Weights (Method 12),” ISO/IEC JTC1/SC29/WG11, m53399, Apr. 2020.

[3] W. Bailer, et al, “Evaluation Framework for Compression of neural networks for multimedia content description and analysis,” ISO/IEC JTC1/SC29/WG11, N18575, July. 2019.