Inter-Layer Kernel Prediction: 프레임 간 Prediction에 기반한 컨볼루션 신경망 가중치 공유 및 모델 압축 방법

이강호, *배성호
경희대학교
ho7719@khu.ac.kr, *shbae@khu.ac.kr

Inter-Layer Kernel Prediction: Weight Sharing and Model Compression of Convolutional Neural Networks Motivated by Inter-frame Prediction

Kang-Ho Lee *Sung-Ho Bae
Kyung Hee University

요 약

본 논문에서는 최근 대두되고 있는 심층신경망 압축 연구에서 가중치 공유와 관련하여 심층신경망 모델 압축방법 *Inter-Layer Kernel Prediction*을 제안한다. 제안 방법은 영상 압축에서 사용되는 프레임 간 prediction 방법을 응용한 컨볼루션 신경망 가중치 공유 및 모델 압축 방법이다. 본 논문은 레이어 간 유사한 kernel들이 존재한다는 것을 발견하고 이를 기반으로 *Inter-Layer Kernel Prediction*을 사용하여 기존 모델 가중치를 보다 더 적은 비트로 표현하여 저장하는 방법을 제안한다. 제안 방법은 CIFAR10/100으로 학습된 ResNet에서 약 4.1 배의 압축률을 달성했으며 CIFAR10으로 학습된 ResNet110에서는 오히려 기존 Baseline 모델에 비해 0.04%의 성능 향상을 기록했다.

1. 서론

최근 심층신경망(Deep Neural Networks, DNN)은 컨볼루션 신경망(Convolutional Neural Networks, CNN)을 필두로 컴퓨터 비전 뿐만 아니라 자연어 처리 등 다양한 분야에서 기존의 전통적인 방법들을 능가하는 성능을 보여주고 있다. 하지만 이러한 성능 향상과 더불어 CNN모델의 크기가 엄청나게 커졌고, 현재 진행중인 연구들도 더 좋은 성능을 위해 더 많은 모델 파라미터를 가지고 크기가 방대해지고 있다.

2000년대 초반에만 해도 구동할 수 없었던 이러한 큰 CNN 모델들은 하드웨어의 발전과 함께 동작이 가능해졌다. 하지만, 모바일 환경과 네비게이션, 키오스크와 같은 임베디드 시스템 환경처럼 작은 저장공간, 비교적 느린 계산 성능 등 컴퓨팅 자원이 제한된 환경에서는 여전히 사용하기 어렵다.

이런 문제를 해결하기 위해 CNN 모델 크기를 줄이거나, 효율적인 CNN구조를 설계하는 방법들이 최근 심층신경망 연구에서 핵심분야로 떠올랐다. 대표적인 방법들로 CNN 모델 압축방법에는 가지치기(Pruning)[1], 양자화(Quantization)[2], 지식 증류(Distillation)[3], 가중치 공유(Weight Sharing)[4]가 있고, 효율적인 CNN모델 구조 설계 방법으로 Depthwise Separable Convolution[5] 등이 있다. 이 방법들은 실질적으로 CNN 모델의 크기를 압축하는데 널리 사용되고 있다.

그 중 가중치 공유 방법은 CNN모델 내의 유사한 가중치를

[1] 본 논문은 과학기술정보통신부 및 정보통신산업진흥원의 '고성능 컴퓨팅 지원' 사업 및 2020 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2018R1C1B3008159)

찾아서 하나의 가중치로 표현하는 방법과 처음부터 구조적으로 같은 가중치를 사용하는 방법이 있다. 두 방법 모두 CNN 파라미터 수를 앞선 논문을 통해 획기적으로 줄인다는 것이 확인되었다. 그리고 두 방법 중 유사한 가중치를 찾아서 하나의 가중치로 표현하는 방법은 H.264[6], H.265[7] 등 동영상 압축에서 주요 모듈로 사용되는 Prediction 방법과 유사한 동기를 가진다.

Prediction 방법은 동영상은 연속적인 프레임에서 프레임 사이뿐만 아니라 각 프레임 내에서 유사한 부분이 많다는 것을 기반으로 한다. 각 프레임을 매크로 블록으로 나누어 이전 프레임 또는 해당 프레임 내에서 유사한 블록들을 찾아서 두 블록 간의 차분을 저장하고 추가적인 기법들과 함께 압축률을 높인다. 본 논문은 프레임 간 Prediction 방법에서 착안하여 가중치 공유를 통한 CNN 모델 압축 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2절에서는 제안하는 방법과 관련된 연구들을 살펴본 후, 3절에서는 본 논문에서 제안하는 방법에 대해 소개한다. 4절에서는 제안 방법에 대한 실험 결과들을 보여주고, 5절에서는 본 논문에 대한 결론 및 향후 연구에 대한 방향을 제시한다.

2. 관련 연구

2.1. 프레임 간 Prediction

프레임 간 Prediction 방법은 H.264, H.265와 같은 블록 단위 움직임 보상 기반의 동영상 압축 표준의 인코딩 단계에서 이산 코사인 변환, 양자화, 엔트로피 부호화와 함께 사용되는 prediction의 한 종류이다. prediction은 intra모드와 inter모드가 있는데 앞서 언급한 프레임 간 prediction은 한 프레임에서 이전 프레임의 블록을 이용하여 해당 프레임의 블록을 예측하는 것이다. Prediction은 동영상이 frame마다 비슷한 부분이 많은 것을 바탕으로 발명되었고, 이는 예측할 블록이 참조할 블록과 서로 유사할 때 높은 압축 성능을 보인다.

2.2. 가중치 공유

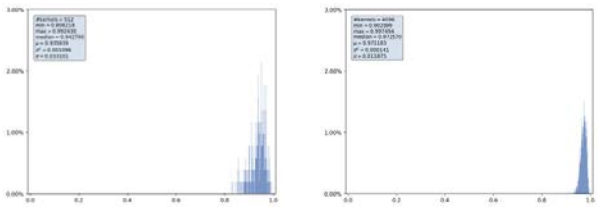
가중치 공유 방법은 CNN모델의 한 레이어에서 유사한 가중치를 찾아서 적은 수의 가중치들로 해당 레이어를 표현하는 방법과 각 레이어마다 가중치를 효율적으로 공유하는 구조를 만드는 방법이 있다.

Deep Compression[8]에서는 각 레이어에서 유사한 가중치를 모아 가중치 공유를 통해 양자화를 하여 CNN 모델을 압축했다. BSConv[9]에서는 레이어 내에 유사한 kernel들이 존재한다는 현상을 확인하고, 일반적인 컨볼루션 레이어에서 각 필터 맵에서 공유되는 하나의 3x3 kernel에 스칼라 곱을 하는 방법을 사용하였다. 최종적으로는 공유하는 방식 대신 수식 재배치를 통해 컨볼루션 레이어를 pointwise convolution 레이어와 depthwise convolution 레이어 순서로 분해했다.

위 연구들에서는 레이어 내의 유사한 가중치들이 존재한다는 것을 기반으로 가중치 공유를 했지만 본 논문은 레이어 간 유사한 kernel들이 존재한다는 것을 발견하고, 프레임 간 prediction 방법에 기반한 CNN 모델 가중치 공유 및 압축 방법, 즉 *Inter-Layer Kernel Prediction*을 제안한다.

3. 제안 방법

본 논문에서는 레이어 간 가중치의 유사도, 즉 레이어 간 kernel의 유사도를 확인하기 위해 각 레이어에서 각 kernel과 이전 레이어들에 위치한 kernel과의 Pearson Correlation Coefficient(PCC)의 절대값이 가장 높은 값들을 히스토그램으로 나타내어 보았다. 이는 그림 1을 보면 확인할 수 있는데, 모두 높은 correlation 값을 가지고 있고, 그림에 없는 다른 레이어들에서도 모두 높은 correlation 값을 가지고 있어 레이어 간 kernel의 유사도가 높은 kernel이 존재함을 알 수 있다.



(a) (b)

그림 1. CIFAR10으로 학습된 ResNet14[10]에서 PCC절대값의 최대값 히스토그램. (a)는 5번 레이어, (b)는 12번 레이어에 대한 히스토그램을 나타냄.

이를 기반으로 본 논문에서는 프레임 간 prediction 방법을 응용하여, *Inter-Layer Kernel Prediction*을 제안한다. *Inter-Layer Kernel Prediction*은 그림 2에서 나타나 있듯이, reference 레이어의 kernel들로 이후 레이어들의 kernel들을 scale factor와 offset factor를 이용하여 만들어 가중치를 공유하여 모델을 압축하는 방법이다. 먼저 reference 레이어를 0번 레이어로 정하고, 이후 레이어들에서 각 kernel들과 가장

유사한 kernel을 reference 레이어에서 찾아 reference 레이어 kernel의 index와 함께 scale factor, offset factor를 계산하여 저장한다. 이때, index는 reference 레이어의 kernel 개수에 맞게 $\lceil \log_2 n \rceil$ bit로 저장하고, 유사도는 PCC의 절대값으로 판단한다. 여기서 Index를 찾는 것을 수식으로 나타내면,

$$\arg\max_{k} \left| r_{K_{0,k}K_{i,j}} \right| \quad (1)$$

이다. 여기서 $r_{K_{0,k}K_{i,j}}$는 kernel $K_{0,k}$와 kernel $K_{i,j}$ 사이의 PCC를 말하고, $K_{0,k}$는 0번 레이어의 k번째 kernel을 말하고, $K_{i,j}$은 i번 레이어의 j번째 kernel을 말한다.

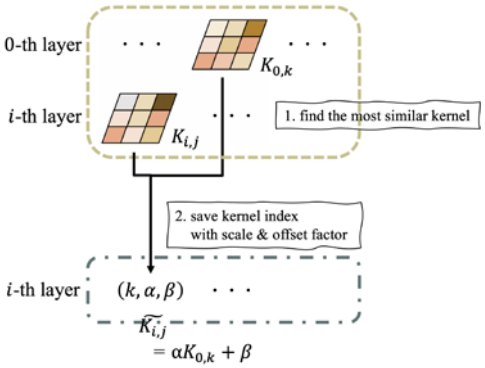


그림 2. 제안 방법 도시. 위의 파선은 CNN 모델을 나타내고, 아래의 1점 쇄선은 메모리 공간을 나타냄.

한편, reference 레이어의 kernel에서 이후 레이어의 kernel을 표현하기 위한 scale factor, offset factor는 correlation 값이 높은 두 kernel 사이의 선형성을 바탕으로 least squared method을 이용하여 구한다. 이는 아래의 식으로 표현할 수 있다.

$$\begin{cases} \alpha = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\ \beta = \bar{y} - \alpha\bar{x} \end{cases} \quad (2)$$

이때 α는 scale factor, β는 offset factor이고, x는 reference 레이어의 kernel, y는 이후 특정 레이어의 kernel을 말한다. 또한, x_i, y_i는 각각 kernel을 일렬로 reshape했을 때, i번째 가중치를 말하고, $\bar{x}$, $\bar{y}$는 각 kernel의 가중치의 평균을 말한다.

최종적으로 추론시간에는 저장된 index, scale factor와 offset factor를 사용하여 reference 레이어 이후 레이어들의 kernel을 복원하여 사용한다. 하지만 kernel 간 유사성이 아무리 크더라도 scale factor와 offset factor로 *Inter-Layer Kernel Prediction*으로 표현된 kernel과 원래 kernel과의 오차가 있으므로 성능 하락을 야기한다. 따라서 본 논문에서는 제안 방법을 CNN 모델에 적용한 후 *Inter-Layer Kernel Prediction*한 가중치들을 초기값으로 하여 fine-tuning을 진행한다. Fine-tuning 시, 학습되는 매 epoch마다 제안 방법을 적용시켜 최종 학습된 가중치가 *Inter-Layer Kernel Prediction*에 최적화된 가중치로 만든다.

4. 실험 결과

실험에 사용한 프레임워크는 PyTorch이고, ResNet20/32/44/56/110 모델을 사용하여 CIFAR10/100 데이터셋으로 학습하였다. 배치 크기는 256, epoch 수는 200, 초기 학습률은 0.1로 설정하였고, 매 epoch 마다 0.98을 곱하여 감소시키면서 학습시켰다. Fine-tuning 시에는 초기 학습률을 0.01로 설정하고 매 epoch 마다 0.98을 곱하여 감소시키면서 학습시켰다. 또한, optimizer는 SGD with Nesterov momentum을 사용하고, momentum은 0.9로 설정하였다.

정확한 실험을 위해 5번의 학습과 함께 fine-tuning을 하여 평균을 낸 결과를 나타내었다. 여기서 압축률은 컨볼루션 레이어의 가중치 bit로만 비교하여 나타냈다. 또한, CIFAR 데이터셋 기반 ResNet의 경우 첫 번째 레이어의 컨볼루션 kernel이 3x16=48개가 존재하므로 index는 6bit를 사용하여 저장했다.

데이터셋	모델	Top-1 accuracy		압축률
		Baseline	Ours	
CIFAR10	ResNet20	92.27	**91.25**±0.16	4.094
	ResNet32	92.66	**92.19**±0.17	4.102
	ResNet44	93.24	**93.13**±0.19	4.106
	ResNet56	93.52	**93.24**±0.19	4.108
	ResNet110	93.73	**93.77**±0.18	4.111
CIFAR100	ResNet20	66.54	**65.70**±0.30	4.094
	ResNet32	68.96	**67.70**±0.11	4.102
	ResNet44	69.57	**68.15**±0.10	4.106
	ResNet56	70.62	**69.41**±0.29	4.108
	ResNet110	72.85	**71.47**±0.18	4.111

표 1. Baseline 대비 제안 방법의 정확도 및 압축률

실험에 사용한 두 데이터셋에서 모두 제안 방법을 적용했을

때, 모든 ResNet에서 약 4.1 배의 압축률을 보여주었고, 압축률 대비 무시할만한 성능 하락을 보여주고 있다. 특히, 큰 모델인 ResNet110에서는 CIFAR10으로 학습된 모델에서는 오히려 0.04%의 성능 상승과 함께 4.11 배의 압축률을 얻었다. 이는 표 1에서 확인할 수 있다.

5. 결론 및 향후 연구

본 논문에서는 레이어 간 유사한 kernel들이 존재한다는 것을 발견하고 동영상 압축에서 사용되는 프레임 간 prediction 방법을 응용하여 가중치 공유를 통한 모델 압축 방법인 *Inter-Layer Kernel Prediction*을 제안하였다. 이는 기존 모델보다 더 적은 bit를 사용하여 저장하는 메모리 효율적인 저장방식이다. 이를 통해 본 논문에서는 기존에 CNN 모델 압축에서 사용되고 있는 양자화 등을 사용하지 않아도 무시할만한 성능 저하로 약 4.1배의 압축률을 얻을 수 있었다.

추후에 ImageNet 실험 등 더 큰 데이터셋을 포함해 다양한 데이터셋에서도 적용시켜 볼 것이다. 또한, reference 레이어를 변경하거나 다양한 실험을 통해 추가적인 evidence를 얻어 압축률을 더 높이거나 압축률은 유지하면서 성능을 더 높일 수 있을 것이다.

참 고 문 헌

[1] Lin, Tao, et al. "Dynamic model pruning with feedback." arXiv preprint arXiv:2006.07253 (2020).

[2] Esser, Steven K., et al. "Learned step size quantization." arXiv preprint arXiv:1902.08153 (2019).

[3] Heo, Byeongho, et al. "A comprehensive overhaul of feature distillation." Proceedings of the IEEE International Conference on Computer Vision. 2019.

[4] Shin, Eunseop, and Sung-Ho Bae. "Global Weight: Network Level Weight Sharing for Compression of Deep Neural Network." Proceedings of the Korean Society of Broadcast Engineers Conference. The Korean Institute of Broadcast and Media Engineers, 2020.

[5] Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arXiv preprint arXiv:1704.04861 (2017).

[6] Wiegand, Thomas, et al. "Overview of the H. 264/AVC video coding standard." IEEE Transactions on circuits and systems for video technology 13.7 (2003): 560-576.

[7] Sullivan, Gary J., et al. "Overview of the high efficiency video coding (HEVC) standard." IEEE Transactions on circuits and systems for video technology 22.12 (2012): 1649-1668.

[8] Han, Song, Huizi Mao, and William J. Dally. "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding." arXiv preprint arXiv:1510.00149 (2015).

[9] Haase, Daniel, and Manuel Amthor. "Rethinking Depthwise Separable Convolutions: How Intra-Kernel Correlations Lead to Improved MobileNets." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[10] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.