JKIICE

Short Paper

한국정보통신학회논문지 Vol. 24, No. 1: 141~144, Jan. 2020

싱글숏 멀티박스 검출기에서 객체 검출을 위한 가속 회로 인지형 가지치기 기반 합성곱 신경망 기법

Convolutional Neural Network Based on Accelerator-Aware Pruning for Object Detection in Single-Shot Multibox Detector

Hyeong-Ju Kang*

*Associate Professor, School of Computer Science and Engineering, Korea University of Technology and Education, Cheonan, 31253 Korea

ABSTRACT

Convolutional neural networks (CNNs) show high performance in computer vision tasks including object detection, but a lot of weight storage and computation is required. In this paper, a pruning scheme is applied to CNNs for object detection, which can remove much amount of weights with a negligible performance degradation. Contrary to the previous ones, the pruning scheme applied in this paper considers the base accelerator architecture. With the consideration, the pruned CNNs can be efficiently performed on an ASIC or FPGA accelerator. Even with the constrained pruning, the resulting CNN shows a negligible degradation of detection performance, less-than-1% point degradation of mAP on VOD0712 test set. With the proposed scheme, CNNs can be applied to objection dtection efficiently.

Keywords : Convolutional neural networks, Pruning, Object detection, Single-shot multibox detector

Ⅰ. 서 론

최근 몇 년 동안 합성곱 신경망(convolutional neural network, CNN)은 객체 검출과 같은 영상 인식 분야에서 탁월한 성능을 보여 주고 있다[1,2]. 객체 검출은 주어진 입력 이미지에서 객체를 검출하여 경계 박스(bounding box)를 찾아내는 작업이며 자율 주행과 같은 분야의 핵심 기술 중 하나이다. 합성곱 신경망이 이 분야에서 높은 성능을 보여 주고 있으나, 많은 양의 저장 공간과 연산을 요구해서 임베디드 환경에서 사용하기에는 어려움이 많다.

저장 공간과 연산의 양을 줄이기 위해 다양한 방법이 제안되어 왔으며, 이 중 가지치기(pruning) 기법은 합성곱 신경망의 일부 계수들을 0이 되게 한다 [3-8]. 0과의 곱셈과 덧셈은 의미가 없으므로 0이 된 계수들은 저장하거나 계산할 필요가 없어져서, 요구되는 저장 공간과 연산의 양을 줄일 수 있다. 그러나 이러한 감소가 실제 연산 시간의 단축으로 그대로 이어지는 것은 아니다. 0이 된 계수들의 분포가 불규칙해서 기반 하드웨어의 성능을 제대로 활용하지 못하게 된다. 이러한 단점을 극복하기 위해 기반 하드웨어의 구조를 고려하여 일정한 패턴을 가지는 가지치기 기법들이 제안되어 왔으며, 이들을 구조적(structured) 가지치기 기법이라고 부른다. [4-8]

그래픽 처리 장치(graphic processing unit, GPU)나 단일 명령 다중 데이터 처리(SIMD)를 고려한 가지치기는 연구되어 왔으나[6,7] ASIC이나 FPGA에서의 가속 회로는 거의 고려되지 않았다. 가속 회로 구조를 고려한 가지치기는 영상 분류 부분에 대해 논문 [8]에서 제안되었다. 그러나 실제 산업계에서 중요한 것은 객체 검출이므로 객체 검출을 위한 합성곱 신경망에도 가속 회로 인지형 가지치기가 적용될 수 있는지 파악해야 할 것이다.

이 논문에서는 가속 회로 인지형 가지치기를 객체 검출 합성곱 신경망에 적용하여 객체 검출에서도 여전히 효과적임을 보였다. 다양한 설정으로 가지치기를 행한 결과 1% 포인트 이하의 성능 저하에서 75%의 계수를 제거할 수 있었다.

이 논문의 구성은 다음과 같다. 2장에서는 객체 검출과 합성곱 신경망을 설명하고, 3장에서는 신경망에서의

Received 20 November 2019, **Revised** 15 December 2019, **Accepted** 28 December 2019

* **Corresponding Author** Hyeong-Ju Kang(E-mail:hjkang@koreatech.ac.kr, Tel:+82-41-560-1420)
Associate Professor, School of Computer Science and Engineering, Korea University of Technology and Education, Cheonan, 31253 Korea

Open Access http://doi.org/10.6109/jkiice.2020.24.1.141 print ISSN: 2234-4772 online ISSN: 2288-4165

가지치기 기법을 소개한다. 4장에서 가속 회로 인지형 가지치기를 제안하고, 5장에서 객체 검출 합성곱 신경망에 적용한 결과를 제시한 뒤, 6장에서 결론을 도출한다.

Ⅱ. 객체 검출과 합성곱 신경망

2.1. 합성곱 신경망

합성곱 신경망은 보통 여러 개의 합성곱 층과 수 개의 완전 연결(fully-connected) 층으로 이루어져 있다. 그리고 그 층들 사이에는 풀링(pooling) 층이나 배치 정규화(batch normalization) 층, 활성화(activation) 층을 넣기도 한다. 대부분의 연산양은 합성곱 층이 차지하고, 대부분의 계수 저장 공간은 완전 연결 층이 요구한다고 알려져 있다. 그래서 현대의 합성곱 신경망들은 완전 연결 층을 하나만 가지거나 사용하지 않는다.

2.2. 객체 검출

객체 검출에서는 주어진 입력 이미지에 있는 객체들을 찾아서 경계 박스를 이용하여 위치 및 크기를 표시한다. 현대의 객체 검출 알고리즘들은 보통 기반 합성곱 신경망과 후처리 과정으로 구성되어 있다. VGG[1]와 같은 합성곱 신경망들이 기반 합성곱 신경망으로 많이 이용된다. 여러 가지의 후처리 과정이 제안되어 왔으며, 많이 이용되는 것 중의 하나가 싱글숏 멀티박스 검출기(single-shot multibox detector, SSD) 방식이다[2]. 이 방식은 성능이 나쁘지 않으면서도 연산양이 적은 것이 특징이다. 싱글숏 멀티박스 검출기에서는 기반이 되는 CNN으로 영상을 처리한 뒤 여러 층을 통해 데이터의 가로, 세로 크기를 줄여나가고, 이들 층의 출력들을 모아서 객체가 있는지 검출한다.

Ⅲ. 신경망의 가지치기

신경망에서의 가지치기는 중요하지 않은 계수들을 0으로 만들어서 저장 공간이나 연산의 양을 줄이는 기법이다[3]. 논문 [3]의 기법에서는 가지치기를 통해 많은 양의 계수들을 없앨 수 있으나 0이 되는 계수들의 패턴이 불규칙한데 반해, 논문 [4-8]에서는 0이 되는 계수들의 규칙성을 고려하는 기법들이 제안되었다. 이 기법들은 채널별, 필터별, 모양별 가지치기로 구별할 수 있다. 예를 들어, 채널별 가지치기에서는 한 채널에 해당하는 계수들을 모두 0으로 만들지를 결정한다. 이러한 가지치기 기법들은 구조적 가지치기라고 부르며, 이와 반대로 0이 되는 계수들의 패턴을 고려하지 않는 방법은 비구조적 가지치기라고 부른다.

Ⅳ. 가속 회로 인지형 가지치기

가속 회로 인지형 가지치기에서는, ASIC이나 FPGA 가속 회로의 처리 형태를 고려하여 0이 될 계수들을 선택한다. 가속 회로에서는 각 층에서, 여러 개의 입력 데이터와 계수들을 함께 가져와서 곱셈을 수행한다. 이전의 가지치기 기법에서는 데이터를 가져오는 경계를 고려하지 않아서, 가지치기된 계수들을 처리할 때 여러 가지 비효율적인 동작이 발생한다.

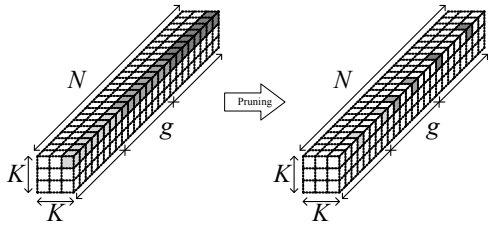

Fig. 1 Accelerator-aware pruning

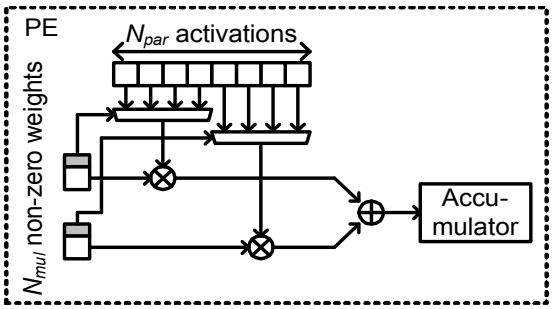

Fig. 2 Accelerator architecture example [8]

곱셈기들을 완전히 사용하려면 곱셈기의 개수와 같은 개수의 0이 아닌 계수들을 가져와야하는데, 이들을 가져왔을 때 그들에 해당하는 입력 데이터들이 한 번에 가져 올 수 있는 경계 안에 속해 있을 것이 보장되지 않는다. 이들 계수들을 처리하기 위해 여러 그룹의 입력

데이터를 가져와야 할 수도 있는 것이다. 그리고 가속 회로가 여러 개의 처리 블록(processing element, PE)들로 구성되어 있다면, 처리 블록들이 같은 개수의 0이 아닌 계수들을 처리한다는 보장도 없다. 처리할 계수의 개수가 적은 처리 블록들은 다른 블록들이 작업을 마치는 것을 기다려야 할 수도 있다.

가속 회로 인지형 가지치기는 가속 회로의 구조를 고려하며 가지치기를 수행해서 이러한 비효율성을 제거할 수 있다. 가속 회로에서 여러 측면을 고려할 수 있으나 이 논문에서는 데이터와 계수를 가져 오는 경계를 고려하였다. 각각의 한 번에 가져오는 입력 데이터 그룹에 대해, 그 그룹에 해당하는 계수들은 항상 같은 개수의 계수들이 0이 되어서 0이 아닌 계수들은 같은 개수가 남도록 한다.

그림 1은 8개씩 묶은 계수들에서 6개가 0이 되고 2개가 남도록 가지치기된 예이다. 그림 1은 어느 합성곱 층에서 필터 한 개의 구조이며, N은 입력 채널의 개수, K는 필터의 합성곱 크기이다. 그리고 g는 가속 회로 인지형 가지치기를 행하는 그룹의 크기이다.

그림 2는 가속 회로 처리 블록의 예이다. 한 처리 블록에 2개의 곱셈기가 있고, 2개의 0이 아닌 계수와 8개의 입력 데이터(activation)를 가져와서, 0이 아닌 계수에 해당하는 입력 데이터를 선택한 뒤 곱셈을 수행한다. 이러한 처리 블록을 위해 가속 회로 인지형 가지치기를 수행한다면, 원래의 계수들을 입력 데이터를 가져오는 경계에 맞추어서 8개씩 묶은 뒤, 각 그룹에서 2개의 계수만 남도록 6개의 계수를 0으로 가지치기한다. 이렇게 하면 0이 아닌 계수들의 경계와 입력 데이터의 경계가 서로 맞게 되어서, 입력 데이터를 추가로 가져올 필요가 없게 된다.

Ⅴ. 실험 결과

가속 회로 인지형 가지치기를 통해 가속 회로 동작의 효율성을 높일 수 있음은 논문 [8]에서 이미 증명되어 있으나, 가지치기에서의 제약으로 인해 신경망의 성능이 낮아질 수 있다. 이 장에서는 객체 검출 합성곱 신경망에 가속 회로 인지형 가지치기를 적용하여도 성능이 충분히 유지됨을 보이려 한다. 그룹 내에서 가지치기를 통해 0으로 만들 계수를 고르는 방법은 여러 가지가 있으나 단순하면서도 효과적으로 알려진 방법에 따라 크기가 작은 계수부터 가지치기를 하였다[3,8].

객체 검출 합성곱 신경망은 VGG16 기반의 싱글숏 멀티박스 검출기를 사용하였고, 입력 이미지의 크기는 300x300을 사용하였다. 싱글숏 멀티박스 검출기의 일부 층은 가속 회로에 맞지 않음이 알려져 있어 우선 논문 [9]를 따라서 가속 회로에 알맞도록 구조를 수정하였고 검출 mAP가 77.63%가 되었다. 여기에 본 논문에서 제안하는 가속 회로 인지형 가지치기를 적용한 뒤 재학습을 시행하였다. 사용한 데이터셋은 VOC 데이터셋이며, VOC0712의 테스트 데이터에 대한 mAP 결과가 표 1이다.

Table. 1 VOC0712 test set mAP after accelerator-aware pruning

Pruning		mAP (%)
g	p	
8	5	76.9
8	6	76.7
8	7	74.3
16	11	76.9
16	12	76.4
16	13	75.3
16	14	74.9
16	15	70.4

표 1에서 각 행은 계수들을 g개 씩 그룹으로 묶은 뒤, 각 그룹 안에서 p개의 계수들을 0으로 만드는 경우에 mAP를 측정한 결과이다. 계수들 중 75%를 가지치기한 경우((g,p)가 (8,6) 또는 (16,12)인 경우)에는 원래의 mAP에 비해 크게 악화되지 않음을 알 수 있다. 가지치기를 더 많이 해서 가지치기 비율이 87.5%가 되면((g,p)가 (8,7) 또는 (16.14)인 경우) mAP가 대략 3% 포인트 정도 감소한다. 싱글숏 멀티박스 검출기 기반의 객체 검출 CNN에 대해 가지치기를 적용한 기존의 결과가 없어서 비교하기는 어려우나, 영상 분류 CNN에 대한 기존 가지치기의 결과에서 합성곱 층에서 대략 75% 정도의 계수를 제거할 수 있는 것[3]과 비교하면, 가지치기에 있어서 가속회로를 위한 제한을 두었음에도 비슷한 성능을 낼 수 있다고 볼 수 있다.

Ⅵ. 결 론

이 논문에서는 객체 검출 합성곱 신경망에 가속 회로 인지형 가지치기를 적용하였다. 합성곱 신경망은 객체 검출에서 높은 성능을 보이지만, 요구되는 저장 공간과 연산의 양이 많아서 임베디드 환경에서 사용되기에 어려움이 많다. 계수들의 가지치기를 통해 이를 어느 정도 해결할 수 있으나, 기존의 가지치기 기법들로 처리된 신경망은 ASIC이나 FPGA 가속 회로에 적합하지 않은 문제가 있다. 가속 회로 인지형 가지치기는 합성곱 신경망의 성능을 악화시키지 않으면서 계수와 연산의 양을 줄일 수 있으며, ASIC이나 FPGA 가속 회로에 적합한 신경망을 생성한다. 본 논문에서 객체 검출 합성곱 신경망에 가속 회로 인지형 가지치기를 적용하여 성능 저하 없이도 많은 양의 계수를 제거할 수 있었다. 이러한 가속 회로 인지형 가지치기는 이후 다른 영상 인식 작업을 위한 합성곱 신경망에도 적용될 수 있을 것이다.

ACKNOWLEDGEMENT

This work was also supported by the 2018 Professor Education and Research Promotion Program of KOREATECH and also supported by IDEC (EDA Tool).

REFERENCES

[1] K. Simonyan, and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proceedings of International Conference on Learning Representations*, pp. 1-14, 2015.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” in *Proceedings of European Conference on Computer Vision*, pp. 21-37, 2016.

[3] S. Han, J. Pool, J. Tran, and W. J. Dally, “Learning both weights and connections for efficient neural networks,” in *Proceedings of Advances in Neural Information Processing Systems*, pp. 1135-1143, 2015.

[4] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, “Learning structured sparsity in deep neural networks,” in *Proceedings of Advances in Neural Information Processing Systems*, pp. 2074-2082, 2016.

[5] Y. He, X. Zhang, and J. Sun, “Channel pruning for accelerating very deep neural networks,” in *Proceedings of International Conference on Computer Vision*, pp. 1398-1406, 2017.

[6] V. Lebedev, and V. Lempitsky, “Fast ConvNets using group-wise brain damage,” in *Proceedings of Computer Vision and Pattern Recognition*, pp. 2554-2564, 2016.

[7] J. Yu, A. Lukefahr, D. Palframan, G. Dasika, R. Das, and S. Mahlke, “Scalpel: Customizing DNN pruning to the underlying hardware parallelism,” in *Proceedings of International Symposium on Computer Architecture*, pp. 548-560, 2017.

[8] H.-J. Kang, “Accelerator-aware pruning for convolutional neural networks,” *IEEE Transactions on Circuits and Systems for Video Technology*, in press.

[9] Y. Ma, T. Zheng, Y. Cao, S. Vrudhula, and J. Seo, “Algorithm-Hardware co-design of single shot detector for fast object detection on FPGAs,” in *Proceedings of International Conference on Computer-Aided Design*, pp. 1-8, 2018.