채널간 압축과 해제를 통한 MobileNetV2 최적화

박진호, 김원준
건국대학교
wyzkssm@gmail.com, wonjkim@konkuk.ac.kr

Further Optimize MobileNetV2 with Channel-wise Squeeze and Excitation

Jinho Park, Wonjun Kim
Konkuk University

Abstract

Depth-wise separable convolution 은 컴퓨터 자원이 제한된 환경에서 기존의 standard convolution을 대체하는데 강력하고, 효과적인 대안으로 잘 알려져 있다.[1] MobileNetV2 에서는 Inverted residual block을 소개한다. 이는 depth-wise separable convolution으로 인해 생기는 손실, 즉 channel 간의 데이터를 조합해 새로운 feature를 만들어낼 기회를 잃어버릴 때, 이를 depth-wise separable convolution 양단에 point-wise convolution(1x1 convolution)을 사용함으로써 극복해낸 block이다.[1] 하지만 1x1 convolution은 채널 수에 의존적(dependent)인 특징을 갖고 있고, 따라서 결국 네트워크가 깊어지면 깊어질수록 효율적이고(efficient) 가벼운(light weight) 네트워크를 만드는데 병목 현상(bottleneck)을 일으키고 만다. 이 논문에서는 channel-wise squeeze and excitation block(CSE)을 통해 1x1 convolution을 부분적으로 대체하는 방법을 통해 이 병목 현상을 해결한다.

1. Introduction

Depth-wise separable convolution은 AlexNet[2]에서 처음 도입된 group convolution의 한 종류로, group size 가 1인 convolution을 의미한다. Depth-wise separable convolution은 3x3 kernel에 사용될 경우 계산 양이 8 ~ 9 배가 줄어들지만, 정확도에선 큰 손실이 없는 아주 효율적인 convolution이라고 할 수 있다.

하지만 depth-wise convolution의 경우, 기존의 standard convolution 과는 다르게 채널 간의 정보 교환이 일어나지 않는다. 이를 MobileNetV2 에선 1x1 convolution을 사용함으로써 극복해낸다. 그 결과 70~75%에 해당하는 네트워크 파라미터(parameters)들이 1x1 convolution으로부터 오게 되고, 이는 MobileNetV2를 한층 더 최적화(Optimization)하는 데 있어서 새로운 bottleneck이 되고 만다. 이 논문에선 channel-wise squeeze and excitation networks[3] 에서 아이디어를 얻어, 1x1 convolution을 대체하면서, 동시에 더 cost effective 하면서, 채널 간에 정보교환의 역할도 수행해줄 수 있는 channel-wise squeeze and excitation block(CSE)을 소개한다.

2. Discussion and Intuition

MobileNetV2를 더 효율적이고, 더 가볍게 만들기 위해선 다양한 실험을 해 볼 수 있었다. Layer 수를 늘리고 줄여 볼 수 있으며, 각 Layer의 channel 수를 바꿔 볼 수 있으며, 반복되는 block의 repetition을 조정해 볼 수 있고, MobileNetV2의 핵심 block인 Inverted Residual block에서, 가운데 bottleneck이 되는 부분의 expansion 파라미터를 조정해 볼 수도 있다. 이 외에도 activation function, different stride 등을 시도해 보았지만 결론적으론 depth-wise convolution의 중요성을 다시 증명하는 실험이 되었다.

따라서 depth-wise convolution을 사용하면서, 다른 부분에서 최적화를 시도해보는 방법을 택하였다.

실험 결과 MobileNetV2 대부분의 파라미터들(70~75%)이 1x1 convolution으로부터 온다는 것을 알 수 있었고, 특히 모델의 깊이가 깊어져, 채널의 수가 늘어나면 늘어날수록, 1x1 convolution이 전체 파라미터 수에 주는 영향은 커져간다는 것을 파악할 수 있었다. 따라서 1x1을 대체할 수 있는, 새로운 feature를 만들어내는 역할을 하면서, 더 가볍고 효율적인 operation을 찾아보는 쪽으로 연구 방향을 정했다.

다양한 아이디어를 실험해 보던 중, SENet(Squeeze-and-Excitation Networks)의 squeeze and excitation operation에서 영감을 얻어, channel-wise average pooling을 통해 각 pixel 단에서 global information을 얻고, 각 pixel의 중요도에 따라 excitation(scale) 해주는 channel-wise squeeze and excitation block(CSE)을 고안하였다.[3] 따라서 CSE block은 channel을 squeeze 하기 때문에 channel 수에 독립적(independent)이며, channel-wise average pooling을 통해 depth-wise separable이 필요로 하는 channel 간의 정보교환을 돕는 역할을 할 수 있게 된다.

3. Channel-wise Squeeze and Excitation blocks

CSE block 은 Figure 1과 같이 순서대로 squeeze(channel wise avg pooling) → Flatten → FC(fully connected, reduction)→ FC(fully connected, scaling)→Stacking→ Excitation(Expanding)으로 이루어져 있다.

FC(reduction)에서 reduction ratio와 cap에 해당하는 hyper parameters 가 추가된다. reduction ratio는 기존의 squeeze and excitation에 있는 hyper parameter이고, cap은 resolution이 낮은 상태에서 reduction ratio에만 의존해서 과도하게 정보를 압축하는 것을 막아주는 새로운 hyper parameter이다. reduction 이 없을 경우, resolution이 높을 때 과도하게 많은 parameter 이 생성되고, cap이 없으면 resolution이 낮을 때 정보가 지나치게 압축 될 수 있다.

CSE block은 1x1 convolution을 완벽히 대체할 수 없고, Figure 1과 같이 bottleneck 구조에서 depth-wise convolution 뒤에 오는 1x1 convolution을 대체하는 데 사용이 된다. 이는 실험적으로 1x1 convolution을 모두 CSE block 만으로 대체했을 때, 모델의 정확도가 급격히 떨어지는 것을 확인함으로써, CSE block 만 사용했을 땐 충분히

새로운 feature를 만들어내지 못했다는 것을 알 수 있었다.

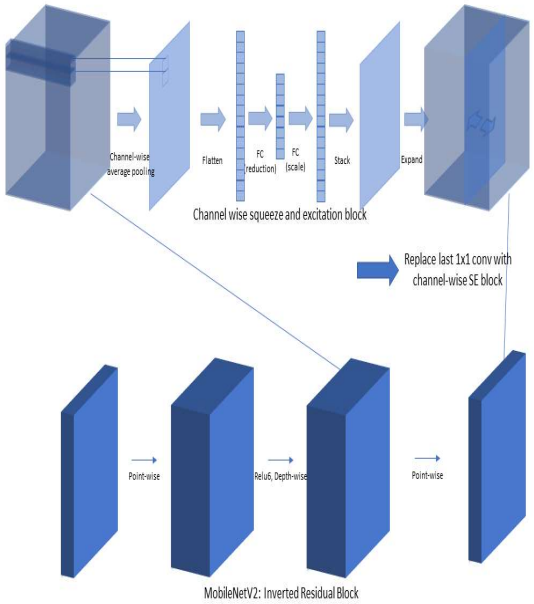

Figure 1. Difference between Inverted Residual block and CSE block

4. Model Architecture

CSE block엔 장단점이 존재하는데, 이는 channel의 수에 독립적인 파라미터 수를 갖지만, resolution에는 비례하는 파라미터 수를 갖는다는 것이다. 1x1 convolution은 이와는 반대로, 채널 수엔 비례하지만, resolution에는 독립적인 파라미터 수를 갖는 특징이 있다.

따라서 Table 1과 같이, 상대적으로 resolution이 크고, 채널 수가 적을 때는 기존의 1x1 convolution을 사용하고, resolution이 줄어들고 채널 수가 늘어났을 때는 CSE block을 사용하는 방식을 사용했다.

Input	Operator	t	c	n	s
32^2 x 3	bottleneck	1	16	1	1
32^2 x 16	bottleneck	6	24	1	1
32^2 x 24	bottleneck	6	32	2	1
32^2 x 32	**bottleneck with CSE**	**2**	**64**	**4**	**2**
16^2 x 64	**bottleneck with CSE**	**2**	**96**	**3**	**2**
8^2 x 96	**bottleneck with CSE**	**2**	**160**	**3**	**1**
8^2 x 160	**bottleneck with CSE**	**2**	**320**	**1**	**2**
4^2 x 320	depthwise conv2d	-	1280	1	1
4^2 x 1280	avgpool 4x4	-	-	1	-

Table 1. Model Architecture of MobileNetV2 with CSE blocks for CIFAR-10.

5. Experiments

5.1. CIFAR-10 image classification

Network	Params(M)	Accuracy(%)
MobileNetV2(baseline)	2.253	91.74
Mix model(3 BTN & 4 CSE), cap 16	**0.363**	**90.03**

Table 2. Comparison between baseline model and mix model on CIFAR-10 image classification.

5.2. CIFAR-100 image classification

Network	Params(M)	Accuracy(%)
MobileNetV2(baseline)	2.253	70.4
MobileNetV2 optimized for CIFAR-100	0.541	63.74
Mix model(3 BTN & 4 CSE), cap 16	**0.478**	**63.34**

Table 3. Comparison between baseline model and mix model on CIFAR-100 image classification.

Image resolution이 작고(32x32), class가 많은 CIFAR-100 같은 경우 baseline model 만큼의 성능을 내긴 어렵지만, CSE block을 사용하지 않은 model과 비교했을 땐 큰 성능 하락 없이 약 12%만큼의 파라미터를 줄일 수 있었다.

5.3. Imagenette image classification

Network	Params(M)	Accuracy(%)
MobileNetV2(baseline)	2.253	91.56
Mix model(3 BTN & 4 CSE), cap 16	**0.444**	**91.31**

* (ImageNette: ImageNet with 10 classes)

Table 4. Comparison between baseline model and mix model on ImageNette image classification.

5.4 Ablation Study

Network	Params(M)	Accuracy(%)
MobileNetV2(baseline)	2.253	91.74
MobileNetV2 optimized for CIFAR10	0.425	90.01
CSE block only, cap 8	0.228	88.13
Mix model(3 BTN & 4 CSE), cap 16	**0.363**	**90.03**

Table 5. CIFAR-10 experiment result

Mix model(기존의 BTN 블락과, BTN블락에서 1x1 convolution을 CSE로 대체한 블락을 사용)의 경우 정확도를 잃지 않고 CIFAR10에 optimized 된 MobileNetV2에 비해 약 15%의 파라미터를 줄일 수 있었다.

Network	Params(M)	Accuracy(%)
MobileNetV2(baseline)	2.253	91.56
CSE block only, cap 8	0.453	85.554
Mix model(3 BTN & 4 CSE), cap 16	**0.444**	**91.31**

Table 6. ImageNette experiment result

CIFAR-10 결과와 마찬가지로, 정확도를 크게 잃지 않고 성공적으로 파라미터를 줄일 수 있었다.

6. Conclusion

이 논문에선 bottleneck인 1x1 convolution을 대체할 수 있는 가볍고 효율적인 CSE block을 제안했다. 하지만 이 CSE block은 항상 1x1 convolution을 대체할 수 있는 것은 아니기에, trade-off를 알고 사용한다면 MobileNetV2을 target dataset에 맞게 더 최적화할 수 있었다.

CIFAR-10 dataset의 경우, Table 5의 MobileNetV2 optimized for CIFAR-10 보다 정확도를 잃지 않고 파라미터 수를 15%가량 줄일 수 있었다. CIFAR-100 dataset의 경우, Table 5의 MobileNetV2 optimized for CIFAR-100 보다 정확도를 잃지 않고 파라미터 수를 12%가량 줄일 수 있었다. ImageNette dataset의 경우, baseline MobileNet2의 20% 에 해당하는 네트워크 크기로 비슷한 성능을 낼 수 있었다.

감사의 글

본 연구는 2021년도 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 결과로 수행되었음 (No.2018-0-00213, SW중심대학(건국대학교)).

References

[1] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E.

Hinton. Imagenet classification with deep convolutional neural networks.

[3] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks.