인간 자세 추정을 위한 경량화 딥러닝 알고리즘 개발

노원준, 문기렴, 한동현, 이병대
경기대학교 컴퓨터공학부
{knoh97, jij7401, mpolio2, blee}@kyonggi.ac.kr

Lightweight Human Pose Estimation Using Deep Learning

Won-Jun Noh, Ki-Ryum Moon, Dong-Hyun Han, Byoung-Dai Lee
Kyonggi Univ.

요 약

사람의 자세 추정은 컴퓨터 비전 분야의 주요 문제 중 하나이다. 최근에는 딥러닝 기술을 활용하여 상당한 성능 개선을 달성하였으나, 계산량이 증가하여 실시간으로 사용하기에는 제약이 많은 경우가 빈번하다. 스마트폰과 같은 모바일 기기에 적용하기 위해서는 자세 정확도를 유지함과 동시에 추정 시간을 감소시켜야 한다. 본 논문에서는 자세 추정 분야에서 우수한 성능을 선보인 Soft Gated Skip Connection을 사용하여 실시간 자세 추정을 가능하도록 하는 개선 방법을 연구하였다. 표준 컨볼루션을 깊이 분리별 컨볼루션으로 대체하고 각 네트워크의 디코더 부분에서 여러 개의 팽창(Dilation)값이 적용된 컨볼루션을 사용하여 합하는 다중 팽창 컨볼루션을 사용하였다. 위 방법으로 정확도는 87%를 달성하였고 추정 시간은 3.6배 감소시켰다. 또한 인코더에서 입력을 처리하기 전에 채널 분할(Channel Split)을 적용하여 기존 채널의 절반을 사용하여 컨볼루션을 처리하고 채널을 합치는 방식으로 사용하였고, 자세 추정 시간은 4.1배 감소하였으며 87%의 정확도를 기록하였다.

Ⅰ. 서 론

인간 자세 추정 알고리즘은 인간의 행동을 감지하기 위한 과제이며 컴퓨터 비전의 주요 문제 중 하나이다. 최근 딥러닝 기술을 집약한 다양한 자세 추정 알고리즘과 모델이 발표되고 있으며, 정확도 또한 비약적으로 개선되었다. 이러한 자세 추정 모델을 이용하여 홈 트레이닝이나 운동 자세 교정 등 사람의 자세를 추정하고 표준 자세와 비교하여 유사도를 측정하는 프로그램이 많이 개발되었다. 최근에는 자세 추정 모델을 서버에서 처리하는 것이 아닌 모바일 기기 자체에서 처리하여 서버로 데이터 전송 없이 자세를 추정하는 엣지 컴퓨팅 기술을 사용한 프로그램이 등장하였다. 하지만 모바일 기기의 성능은 그래픽 카드를 사용하는 컴퓨터에 비해 매우 떨어지고, 일반적인 자세 추정 모델은 정확도를 높이기 위해 신경망이 깊고 지연 시간이 길어 모바일 기기에서의 실시간 자세 추정에 사용하기에는 어렵다.

본 논문에서는 자세 추정에 대해 정확도가 높은 Soft Gated Skip Connection(SGSC)[6]를 사용하여 실시간 자세 추정을 가능하도록 하는 개선 방법을 연구하였다. 일정 수준의 정확도를 유지하면서 동시에 실시간으로 사용할 수 있는 딥러닝 기반의 경량 자세 추정 알고리즘을 제안한다.

Ⅱ. 본론

1. 경량화 대상 모델 소개

본 논문에서는 Stacked Hourglass Network(SHG)를 기반으로 한 SGSC 모델을 이용하여 자세 추정을 실시간으로 수행하기 위해 경량화 알고리즘 연구를 수행하였다. SHG는 컨볼루션을 통한 잔차 모듈과 최대 풀링 레이어를 사용하여 특징을 추출하고 저차원으로 다운 샘플링(downsampling)한다. 가장 낮은 해상도에 도달한 후 업 샘플링(upsampling)을 수행하면서 각 해상도별 특징을 조합하는 동작을 여러 번 반복하는 네트워크이다. 인코더-디코더 구조로 이루어진 네트워크는 대칭적인 구조를 사용한다. SGSC는 SHG의 잔차 블록을 3번의 컨볼루션을 수행하는 Soft Gated Block(SGB)으로 변경한다. 디코더와 인코더를 연결하는 스킵 커넥션에 각 채널 별 역전파를 통해 학습된 가중치를 입력값에 적용하여 특징 간 중요도를 조절한다. 각 컨볼루션이 수행된 후 텐서를 채널별 차원으로 연결하여 최종 관절 위치를 생성한다.

1-1. 인코더 - 깊이 분리별 컨볼루션과 팽창 컨볼루션

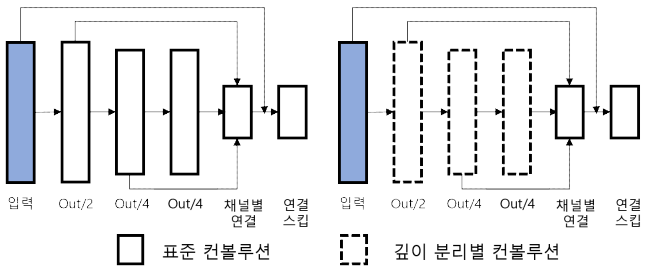
일반적인 컨볼루션에서 특징 맵 한 개를 생성하는데 사용하는 가중치 수는 커널의 수와 크기에 비례한다. 본 논문에서는 SGSC의 인코더 부분의 SGB를 MobileNet[4]의 깊이 분리별 컨볼루션(DS)으로 교체하는 방법을 제안한다 (그림 1). 입력 텐서를 채널 별로 분리하여 컨볼루션을 수행하고, 1x1 컨볼루션을 이용하여 분리된 채널를 통합한다. 깊이 컨볼루션을 수행할 때 표준 컨볼루션보다 적은 가중치를 사용하여 동일한 크기의 텐서를 도출한다.

또한 SGB에서 깊이 분리별 컨볼루션을 수행할 때 필터 내부에 제로 패딩(zero-padding)을 추가하여 수용 범위(receptive Field)를 늘리는 팽창된 컨볼루션(Dilated Convolution)을 사용하였다. 표준 컨볼루션과 같은 가중치를 사용하면서 수용 범위를 넓게 인식하여 특징 맵을 추출한다. 이 연구에서는 팽창 값을 2로 설정하여 수용 범위를 5x5로 사용하였다.

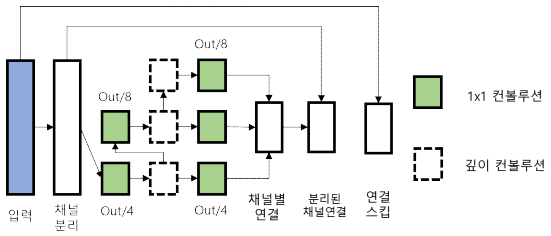
1-2. 인코더 - 채널 분할, 채널 셔플

SGB에서 표준 컨볼루션을 대체하여 깊이 분리별 컨볼루션을 사용하여 계산량를 대폭 감소시켰지만, 모델의 메모리 접근 횟수는 무조건적으로 감소하지 않는다. 컨볼루션을 그룹으로 통합하여 수행하면 그룹이 커질수록 메모리 접근 횟수는 증가한다. 이를 해결하기 위해 1x1 컨볼루션에 그룹을 적용하지 않는 대신, 특징 맵을 절반으로 나누어 사용한다. 또한 요소별 덧셈을 사용하지 않고 연결을 수행한다 (그림 2).

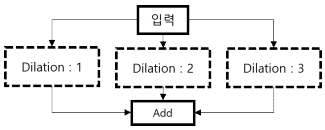
채널 분할과 셔플을 SGB에 적용하기 위해 첫 입력값을 채널 기준으로 절반 분할 하여 진행한다. 나눠진 채널의 한 부분은 깊이 분리별 컨볼루션을



(그림 1) (a) 일반 컨볼루션을 사용한 SGB (b) 깊이 분리별 컨볼루션을 사용한 SGB



(그림 2) 채널 분리와 채널 셔플을 사용한 SGB



(그림 3) 다중 팽창 컨볼루션

	머리	어깨	팔꿈치	손목	허리	무릎	발목	평균
SGB-DS-MD	96.5	94.1	87.4	81.9	86.8	82.1	77.3	87.2
SGB-SPLIT-MD	95.9	94.5	87.7	81.9	87.2	82.4	77.4	87.4
SGSC[6]	98.6	97	93	89.2	91.7	88.9	86	92.4

(표 1) 각 모델의 키포인트 별 정확도

	가중치 수	메모리 접근 빈도	지연 시간
SGB-DS-MD	2.29M	4.85	280ms
SGB-SPLIT-MD	1.57M	4.35	240ms

(표 2) 각 모델 별 가중치 및 계산량, 지연 시간

수행하고, 다른 한 부분은 컨볼루션을 수행하지 않고 깊이 분리별 컨볼루션을 수행한 채널과 연결한다. 분리된 가중치의 학습을 돕기 위해 연결한 채널에 대해 채널 셔플(Channel Shuffle)을 수행한 뒤 최종 결과물을 도출한다.

1-3. 디코더 - 다중 팽창 블록

표준 컨볼루션에서, 커널의 크기를 증가하여 사용하면 수용 범위가 넓어져 텐서의 전체적인 특징을 원활하게 추출할 수 있다. 하지만 커널의 크기가 증가하면 과적합의 위험이 있고 계산량이 크게 증가한다. 이를 해결하기 위해 LightWeight Stacked Hourglass Network[3]에서 제안한 다중 팽창 블록(MultiDilated Block)을 사용하였다 (그림 3). 팽창 컨볼루션을 하나만 사용한다면 내부의 제로 패딩 크기가 증가하여 특징 맵을 정확히 추출할 수 없다. 따라서 3개의 팽창 컨볼루션을 이용하여 안정적으로 넓은 수용 범위를 사용한다. 1~3까지의 팽창 값이 적용된 깊이 컨볼루션을 처리한 후 가중치별 덧셈을 적용하여 1x1 컨볼루션으로 채널을 통합한다. 다중 팽창 블록은 모래시계 네트워크의 디코더 부분 중 업 샘플링을 처리한 뒤 수행한다. 디코더가 가진 각 해상도 별 특징을 단일 표준 컨볼루션보다 적은 가중치로 넓은 수용 범위를 가지면서 추출한다.

2. 실험 결과

배치 크기는 16, 에폭은 210, 학습률은 2e-3으로 설정하였다. 데이터셋은 2만 개의 관절 키포인트가 라벨링된 MPII 데이터셋[7]을 사용하였다. 정확도 측정 방식은 각 관절의 점수를 0.5를 임계값으로 설정하여 평가하였다. 표 1과 표2에 기술된 SGB-DS-MD는 인코더에 깊이 분리 컨볼루션을, 디코더에 다중 팽창 컨볼루션을 사용하였고, SGB-SPLIT-MD는 인코더에 채널 분리, 셔플을, 디코더는 다중 팽창 컨볼루션을 사용하였다. SGB-SPLIT-MD는 채널 분할을 수행하여 처리해야 하는 가중치가 DS보다 적고 메모리가 감소하면서, 정확도는 소폭 증가하였다. 두 모델은 머리 부분을 제외하고 0.3% 이내의 오차가 존재한다. 지연 시간은 기존 모델과 비교하여 SGB-SPLIT-MD는 3.5배, SGB-SPLIT-MD 모델은 4.1배 감소하였다.

Ⅲ. 결론

본 논문에서는 자세 추정 분야에서 우수한 성능을 선보인 SGSC 모델을 실시간 사용이 가능하게 하기 위한 다양한 기법을 제안하였다. SGB-SPLIT-MD 모델이 SGB-DS-MD보다 성능이 조금 높으면서 적은 가중치를 사용하여 추론 시간이 1.1배 감소하였다. 또한 경량화를 위해 144로 설정한 입력 채널을 성능 향상을 위해 256으로 설정하여 학습한 결과 정확도는 소폭 증가하였지만, 가중치와 메모리 사용량은 3배 증가하였다. 향후 다른 자세 추정 데이터셋인 COCO, LSP를 이용해 유효성 검증을 수행할 예정이다.

ACKNOWLEDGMENT

이 논문은 2020년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2020R1A6A1A03040583)

참 고 문 헌

[1] Newell Alejandro,Yang Kaiyu,Deng Jia. "Stacked Hourglass Networks for Human Pose Estimation" Lecture Notes in Computer Science : 483-499.

[2] Ma Ningning,Zhang Xiangyu,Zheng Hai-Tao,Sun Jian. "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design" Lecture Notes in Computer Science : 122-138.

[3] Kim, S. T., & Lee, H. J. “Lightweight stacked hourglass network for human pose estimation.” Applied Sciences, 10(18), 6497.

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

[5] Zhang, X., Zhou, X., Lin, M., & Sun, J. “Shufflenet: An extremely efficient convolutional neural network for mobile devices.” IEEE conference on computer vision and pattern recognition : pp. 6848-6856

[6] Bulat, Adrian, et al. "Toward fast and accurate human pose estimation via soft-gated skip connections." 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE, 2020.

[7] Andriluka, Mykhaylo, et al. "2d human pose estimation: New benchmark and state of the art analysis." Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. 2014.