인체 자세 추정을 위한 다중 해상도 디컨볼루션 출력망

강원준 조남익
서울대학교 전기정보공학부
kangwj1995@snu.ac.kr, nicho@snu.ac.kr

Multi-Scale Deconvolution Head Network for Human Pose Estimation

Won Jun Kang Nam Ik Cho
Seoul National University

요 약

최근 딥러닝을 이용한 인체 자세 추정(human pose estimation) 연구가 활발히 진행되고 있다. 그 중 구조가 간단하면서도 성능이 강력하여 널리 사용되고 있는 딥러닝 네트워크 모델은 이미지 분류(image classification)에 사용되는 백본 네트워크(backbone network)와 디컨볼루션 출력망(deconvolution head network)을 이어 붙인 구조를 갖는다[1]. 기존의 디컨볼루션 출력망은 디컨볼루션 층을 쌓아 낮은 해상도의 특징맵을 모두 높은 해상도로 변환한 후 최종 인체 자세 추정을 하는데 이는 다양한 해상도에서 얻어낸 특징들을 골고루 활용하기 힘들다는 단점이 있다. 따라서 본 논문에서는 매 디컨볼루션 층 이후에 인체 자세 추정을 하여 다양한 해상도에서 연산을 하고 이를 종합하여 최종 인체 자세 추정을 하는 방법을 제안한다. 실험 결과 Res50 과 기존의 디컨볼루션 출력망의 경우 0.717 AP 를 얻었는데 Res101 과 기존의 디컨볼루션 출력망을 사용한 결과 50% 이상의 파라미터 수 증가와 함께 0.727 AP, 즉 0.010AP 의 성능 향상이 이루어졌다. 이에 반해 Res50 에 다중 해상도 디컨볼루션 출력망을 사용한 결과 약 1%의 파라미터 수 증가 만으로 0.720 AP, 즉 0.003 AP 의 성능 향상이 이루어졌다. 이를 통해 디컨볼루션 출력망 구조를 개선하면 매우 적은 파라미터 수 증가 만으로도 인체 자세 추정의 성능을 효과적으로 향상시킬 수 있음을 확인하였다.

1. 서론

인체 자세 추정은 컴퓨터 비전에 있어서 중요한 연구 분야 중 하나로 손꼽힌다. 인체 자세 추정의 목표는 인간이 등장하는 사진 또는 영상의 입력에 대해 손목, 팔꿈치, 골반 등과 같은 인간의 여러 부위 및 관절의 위치를 추정하는 것이다. 인체 자세 추정의 경우 손목, 팔꿈치 그리고 어깨는 하나의 팔로 이어져 있다는 등의 사전 상호 관계가 존재한다. 즉 장애물에 가린 팔꿈치의 위치를 손목과 어깨의 위치를 이용해 간접적으로 추정하는 것이 가능하다는 점에서 일반적인 객체 검출 분야와 차별화된다. 이러한 인체 자세 추정 기술은 스포츠, 로봇, 애니메이션, 가상현실, 증강현실 등의 다양한 분야에서 응용될 수 있다.

최근 딥러닝의 급격한 발전에 힘입어, 인체 자세 추정 분야 역시 딥러닝 네트워크 모델을 이용한 연구가 강세를 보이고 있다. [2, 3, 4, 5, 6]. 그 중 구조가 간단하면서도 성능이 강력하여 널리 사용되고 있는 딥러닝 네트워크 모델은 이미지 분류 분야에서 사용되는 백본 네트워크와 디컨볼루션 출력망을 이어 붙인 구조를 갖는다[1]. 특징맵(feature map)을 얻기 위해 사용되는 백본 네트워크의 종류에는 ResNet, HRNet 등이

있으며 특히 HRNet 을 이용한 인체 자세 추정 모델은 현재 가장 좋은 성능을 보이는 모델 중 하나이다[7, 8] . 현재 대부분의 인체 자세 추정 연구는 강력한 백본 네트워크 설계에 초점을 맞추고 있고 디컨볼루션 출력망에 대한 연구는 거의 이루어지지 않고 있다.

디컨볼루션 출력망은 백본 네트워크로 얻어진 특징맵의 해상도를 높여 최종 히트맵 출력의 해상도와 같아지게 하기 위해 사용된다. 이를 위해 기존의 디컨볼루션 출력망은 디컨볼루션 층 여러 개를 단순히 이어붙인 구조를 사용하게 되는데 이는 구현이 쉽고 구조가 간단하다는 장점이 있다. 그러나 이러한 방법은 낮은 해상도의 특징맵을 높은 해상도의 특징맵으로 일단 모두 변환한 후 최종 인체 자세 추정을 하기 때문에 다양한 해상도에서 얻어낸 특징들을 골고루 활용하기 힘들다는 단점이 있다.

따라서 본 논문에서는 매 디컨볼루션 층 마다 인체 자세 추정을 하여 다양한 해상도에서 연산을 하고 이를 종합하여 최종 인체 자세 추정을 하는 방법을 제안한다.

2. 다중 해상도 디컨볼루션 출력망

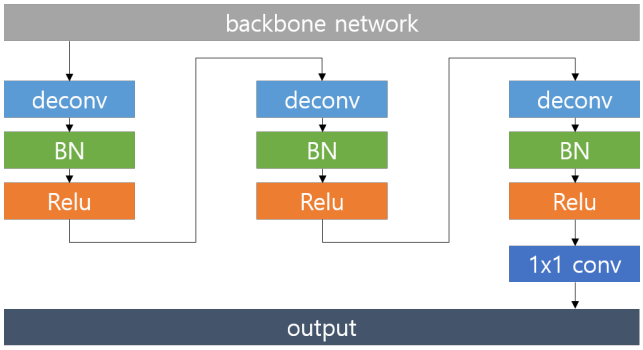

그림 1. 기존의 디컨볼루션 출력망

기존의 디컨볼루션 출력망은 특징맵의 해상도를 키우는 것에 초점을 맞춰서 설계가 되었다. 예를 들어 백본 네트워크로 얻어진 특징맵의 해상도가 (8, 6)이고 최종적으로 얻어야 하는 히트맵 출력의 해상도가 (64, 48)이라면 특징맵의 해상도를 총 8 배 키워야 한다. 이를 위해 출력의 해상도가 입력의 해상도의 2 배인 디컨볼루션 층을 총 3 개 설계하여 나란히 이어 붙이면 기존의 디컨볼루션 출력망이 완성된다. 그런데 이러한 방법은 특징맵의 해상도를 8 배 키운 후의 정보만을 활용하여 1x1 컨볼루션을 통해 히트맵을 추정하기 때문에 특징맵의 해상도를 1 배, 2 배, 4 배 키운 후의 정보를 효과적으로 활용하기 힘들다. 그림 1 은 기존의 디컨볼루션 출력망의 구조를 나타낸다. 특징맵의 해상도를 높이기 위해 3 번의 디컨볼루션 층을 거친 후에 최종 히트맵 출력을 예측하기 위해 1x1 컨볼루션 층을 통과하는 것을 확인할 수 있다.

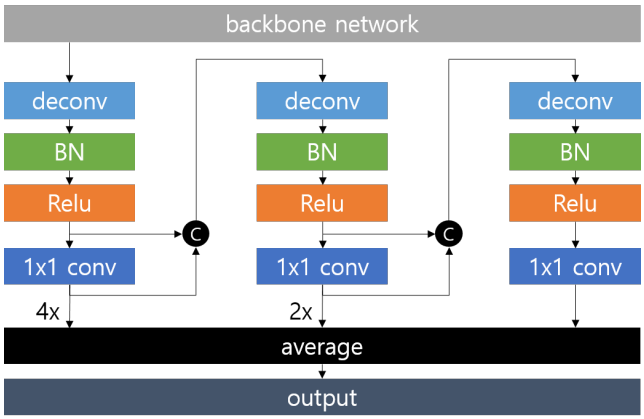

그림 2. 제안한 다중 해상도 디컨볼루션 출력망

기존의 디컨볼루션 출력망이 다양한 해상도의 정보를 활용하지 못한다는 점을 개선하기 위해 본 논문은 다중 해상도 디컨볼루션 출력망을 제안한다. 이를 위해 매 디컨볼루션 층 이후에 1x1 컨볼루션 연산을 추가하였다. 이렇게 얻어진 각각의 1x1 컨볼루션 출력은 최종 히트맵 출력의 크기와 같아지도록 2 의 거듭제곱 배율의 보간을 거친 후 산술 평균을 통해 최종 히트맵 출력을 얻게 된다. 또한 매 1x1 컨볼루션 층의 입력과 출력은 다음 디컨볼루션 층의 입력으로 함께 들어간다. 이는 각각의 디컨볼루션 층에 더욱 풍부한 정보를 전달해주는 역할을 한다. 그림 2 는 총 3 가지 해상도의 정보를 사용하는 다중 해상도 디컨볼루션 출력망의 구조를 나타낸다. 각각의 디컨볼루션 층 출력은 4 배, 2 배, 1 배의 보간이 이루어진 후 산술 평균을 통해 최종 히트맵 출력을 예측한다. 다중 해상도 디컨볼루션 출력망의 손실 함수는 기존의 디컨볼루션 출력망과 마찬가지로 L2 손실 함수를 사용하였다.

3. 실험 방법

본 논문에는 다중 해상도 디컨볼루션 출력망을 이용한 인체 자세 추정 실험을 위해 COCO(Common Objects in Context) 데이터셋을 사용하였다. COCO 데이터셋은 총 약

20 만 장의 사진과 약 25 만 명의 사람이 등장하고 각각의 사람에 대해 17 개의 부위가 표시되어있다. 이 중, 약 57,000 장의 사진과 약 15 만명의 사람이 등장하는 COCO train2017 데이터셋을 이용해 모델을 학습하였다. 성능 확인에는 COCO val2017 데이터셋을 사용하였다.

다중 해상도 디컨볼루션 출력망에 사용할 백본 네트워크로는 ResNet50 을 사용하였다. ResNet50 과 기존의 디컨볼루션 출력망을 사용한 경우, ResNet50 과 다중 해상도 디컨볼루션 출력망을 사용한 경우, ResNet101 과 기존의 디컨볼루션 출력망을 사용한 경우를 각각 실험하여 결과를 비교하였다. 또한 각각의 방법에 사용되는 파라미터 수를 계산하여 비교하였다.

실험에 사용된 다중 해상도 디컨볼루션 출력망의 종류는 총 3 가지이다. ResNet50 으로 얻어진 특징맵 해상도를 8 배 키운 후의 정보만을 사용하는 방법이 기존의 디컨볼루션 출력망이고, 특징맵 해상도를 4 배, 8 배 키운 후의 정보를 사용하는 것이 ResNet50_2 해상도 방법이다. 마찬가지로 ResNet50_3 해상도 방법은 특징맵의 해상도를 2 배, 4 배, 8 배 키운 후의 정보를 사용하고 ResNet50_4 해상도 방법은 특징맵의 해상도를 1 배, 2 배, 4 배, 8 배 키운 후의 정보를 모두 사용한다. 그림 2 는 ResNet50_3 해상도 방법의 구조를 나타낸다.

4. 실험 결과

COCO val2017 데이터셋을 이용해 기존의 방법과 본 논문에서 제안된 방법의 성능을 비교하였다. 성능 지표로는 AP(Average Precision)과 AR(Average Recall)을 사용하였고 실험 결과는 각각 표 1 과 표 2 와 같다.

	AP	AP .5	AP .75	AP (M)	AP (L)
Res50	0.717	0.898	0.794	0.679	0.786
Res50_2 해상도	0.719	0.899	0.796	0.683	0.787
Res50_3 해상도	0.720	0.898	0.797	0.683	0.788
Res50_4 해상도	0.719	0.897	0.794	0.682	0.788
Res101	0.727	0.900	0.807	0.691	0.795

표 1. 제안된 방법과 기존 방법의 AP 비교

	AR	AR .5	AR .75	AR (M)	AR (L)
Res50	0.773	0.937	0.840	0.728	0.839
Res50_2 해상도	0.776	0.936	0.844	0.732	0.839
Res50_3 해상도	0.776	0.938	0.844	0.732	0.839
Res50_4 해상도	0.775	0.937	0.842	0.731	0.839
Res101	0.783	0.939	0.852	0.740	0.845

표 2. 제안된 방법과 기존 방법의 AR 비교

실험 결과, 본 논문에서 제안된 다중 해상도 디컨볼루션 출력망의 경우 3 개의 해상도를 사용한 Res50_3 해상도 방법이 2 개 혹은 4 개의 해상도를 사용한 경우보다 뛰어남을 확인했다. 또한 각 방법의 효율성을 비교하기 위해 Res50_3 해상도 방법의 파라미터 수를 기존 방법들의 파라미터 수와 비교하여 표 3 에 나타내었다.

	파라미터 수
Res50	33999697
Res50_3 해상도	34450316
Res101	52991825

표 3. 각 방법의 파라미터 수 비교

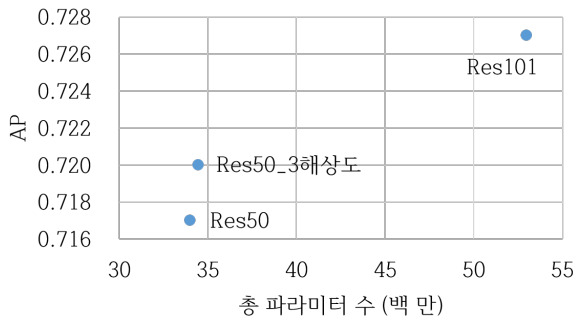

그림 3. 파라미터 수 대비 AP

실험 결과 AP 의 경우 Res50 은 0.717, Res50_3 해상도 방법은 0.720, Res101 은 0.727 의 성능을 얻었다. 또한 AR 의 경우 Res50 은 0.773, Res50_3 해상도 방법은 0.776, Res101 은

0.783 의 성능을 얻었다. 즉 AP 와 AR 모두 Res50 에서 Res101 로 0.010 의 성능 개선이 이루어질 때 Res50 에서 Res50_3 해상도로는 0.003 의 성능 개선이 이루어졌다. 파라미터 수를 비교해보면 Res101 의 경우 Res50 보다 필요한 파라미터 수가 50% 이상 증가했지만 Res50_3 해상도 방법의 경우 Res50 보다 필요한 파라미터 수가 약 1% 증가하는데 그쳤다. 그림 3 은 백본 네트워크를 수정했을 때와 디컨볼루션 출력망을 수정했을 때의 파라미터 수 대비 성능을 그래프로 나타낸 것이다. Res50 과 Res50_3 해상도를 잇는 선의 기울기가 Res50 과 Res101 을 잇는 선의 기울기보다 훨씬 큼을 확인할 수 있다. 즉 백본 네트워크가 아닌 디컨볼루션 출력망을 개선하는 것 만으로도 인체 자세 추정의 효율적인 성능 개선이 이루어질 수 있음을 알 수 있다. 그림 4 는 Res50_3 해상도 방법을 사용하여 사진 안에 등장하는 사람들에 대해 인체 자세 추정을 한 결과의 예시이다.

그림 4. 인체 자세 추정 결과

5. 결론

본 논문에서는 인체 자세 추정을 위한 다중 해상도 디컨볼루션 출력망을 제안하였다. 다중 해상도 디컨볼루션 출력망을 사용한 결과 기존의 디컨볼루션 출력망보다 매우 적은 파라미터 수 증가 만으로 유의미한 성능 향상을 얻을 수 있었다. 이는 성능 향상을 위해 백본 네트워크를 Res50 에서 Res101 으로 바꿀 때 상당한 파라미터 수가 증가하는 것과 대조적이다. 즉 백본 네트워크가 아닌 디컨볼루션 출력망을 개선하는 것 만으로도 인체 자세 추정의 효율적인 성능 향상을 기대할 수 있다는 사실을 확인하였다.

본 논문에서는 백본 네트워크로 ResNet 을 사용하는 경우에 대해서만 실험을 하였지만 다양한 백본 네트워크, 특히 현재 가장 좋은 성능을 보이는 백본 네트워크인 HRNet 에 대해서도 디컨볼루션 출력망 개선을 통한 성능 향상 연구를 기대해볼 수 있을 것이다. 또한 COCO 데이터셋 뿐만 아니라 MPII Human Pose D ataset 등의 다른 인체 자세 추정 데이터셋에 대해서도 실험을 하여 성능을 비교해볼 수 있을 것이다.

감사의 글

이 논문은 FuriosaAI 및 2020 년도 BK21 플러스 창의정보기술 인재양성사업단에 의하여 지원되었음.

참고 문헌

[1] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In ECCV, 2018.

[2] A. Toshev and C. Szegedy. D eeppose: H uman pose estimation via deep neural networks. In CVPR, 2014.

[3] J. J . Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In NIPS, 2014.

[4] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In CVPR, 2015.

[5] S. Wei, V. Ramakrishna, T. Kanade, and Y. S heikh. Convolutional pose machines. In CVPR, 2016.

[6] A. N ewell, K . Yang, and J. D eng. Stacked hourglass networks for human pose estimation. In ECCV, 2016.

[7] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In CVPR, 2019.

[8] He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. In CVPR, 2016.