JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 9, pp. 17-25, September 2023
https://doi.org/10.9708/jksci.2023.28.09.017

Lightening of Human Pose Estimation Algorithm Using MobileViT and Transfer Learning

Kunwoo Kim*, Jonghyun Hong*, Jonghyuk Park*

*Student, Dept. of AI, Big Data & Management, Kookmin University, Seoul, Korea
*Student, Dept. of AI, Big Data & Management, Kookmin University, Seoul, Korea
*Assistant Professor, Dept. of AI, Big Data & Management, Kookmin University, Seoul, Korea

[Abstract]

In this paper, we propose a model that can perform human pose estimation through a MobileViT-based model with fewer parameters and faster estimation. The based model demonstrates lightweight performance through a structure that combines features of convolutional neural networks with features of Vision Transformer. Transformer, which is a major mechanism in this study, has become more influential as its based models perform better than convolutional neural network-based models in the field of computer vision. Similarly, in the field of human pose estimation, Vision Transformer-based ViTPose maintains the best performance in all human pose estimation benchmarks such as COCO, OCHuman, and MPII. However, because Vision Transformer has a heavy model structure with a large number of parameters and requires a relatively large amount of computation, it costs users a lot to train the model. Accordingly, the based model overcame the insufficient Inductive Bias calculation problem, which requires a large amount of computation by Vision Transformer, with Local Representation through a convolutional neural network structure. Finally, the proposed model obtained a mean average precision of 0.694 on the MS COCO benchmark with 3.28 GFLOPs and 9.72 million parameters, which are 1/5 and 1/9 the number compared to ViTPose, respectively.

▸ **Key words**: Deep Learning, Computer Vision, Keypoint Detection, Human Pose Estimation, Vision Transformer

• First Author: Kunwoo Kim, Corresponding Author: Jonghyuk Park
*Kunwoo Kim (kimkunu6632@kookmin.ac.kr), Dept. of AI, Big Data & Management, Kookmin University
*Jonghyun Hong (jody1188@kookmin.ac.kr), Dept. of AI, Big Data & Management, Kookmin University
*Jonghyuk Park (jonghyuk@kookmin.ac.kr), Dept. of AI, Big Data & Management, Kookmin University
• Received: 2023. 07. 19, Revised: 2023. 08. 30, Accepted: 2023. 08. 30.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

본 논문에서는 매개변수가 더 적고, 빠르게 추정 가능한 MobileViT 기반 모델을 통해 사람 자세 추정 과업을 수행할 수 있는 모델을 제안한다. 기반 모델은 합성곱 신경망의 특징과 Vision Transformer의 특징이 결합한 구조를 통해 경량화된 성능을 입증한다. 본 연구에서 주요 매커니즘이 되는 Transformer는 그 기반의 모델들이 컴퓨터 비전 분야에서도 합성곱 신경망 기반의 모델들 대비 더 나은 성능을 보이며, 영향력이 커지게 되었다. 이는 사람 자세 추정 과업에서도 동일한 상황이며, Vision Transformer기반의 ViTPose가 COCO, OCHuman, MPII 등 사람 자세 추정 벤치마크에서 모두 최고 성능을 지키고 있는 것이 그 적절한 예시이다. 하지만 Vision Transformer는 매개변수의 수가 많고 상대적으로 많은 연산량을 요구하는 무거운 모델 구조를 가지고 있기 때문에, 학습에 있어 사용자에게 많은 비용을 야기시킨다. 이에 기반 모델은 Vision Transformer가 많은 계산량을 요구하는 부족한 Inductive Bias 계산 문제를 합성곱 신경망 구조를 통한 Local Representation으로 극복하였다. 최종적으로, 제안 모델은 MS COCO 사람 자세 추정 벤치마크에서 제공하는 Validation Set으로 ViTPose 대비 각각 5분의 1과 9분의 1만큼의 3.28GFLOPs, 972만 매개변수를 나타내었고, 69.4 Mean Average Precision을 달성하여 상대적으로 우수한 성능을 보였다.

▸ **주제어**: 딥러닝, 컴퓨터 비전, 키포인트 탐지, 사람 자세 추정, 비전 트랜스포머

I. Introduction

최근에는 센서를 활용한 고전적인 사람 자세 추정에서 벗어나, 컴퓨터 비전기술 기반 자세 추정 알고리즘들이 개발되고 있다[1]. 딥러닝은 다층 인공 신경망을 통해 Dataset으로부터 특징들을 학습하며, 이를 통해 컴퓨터 비전과 사람 자세 추정 과업에서도 높은 성능을 달성하고 있다[2-4]. 최신 연구에서는 사진 속 이차원 인물 외, 삼차원 인물 및 실시간 인물 추적에도 딥러닝 기반 모델들을 활용하고 있다. 하지만, 딥러닝 기반 모델들은 깊은 신경망 층과 복잡한 알고리즘을 사용하고 있으며, 이러한 방법들은 모델을 더욱 무겁고 복잡한 구조로 바꾸며 여러 제약 사항을 만든다. 예를 들어 Transformer[5]기반의 Vision Transformer(ViT)[6]는 약 8600만 개의 매개변수가 사용된다. 이는 합성곱 신경망 기반의 MobileNet보다 약 10배 많은 수치이다. 무겁고 복잡한 모델을 사용한다는 것은 최적화 문제의 어려움을 발생시키고, 과적합 방지를 위한 광범위한 Data 증강기법과 L2 규제를 요구하며, 다른 알고리즘들과의 비교 및 분석을 더욱 어렵게 만든다[7].

이에 본 연구에서는 사람 자세 추정 분야에서 더 낮은 복잡성과 적은 매개변수를 갖지만, 좋은 성능을 보이는 경량화된 알고리즘을 제시하고자 한다. 기존, 컴퓨터 비전 분야에서 주로 활용되었던 신경망 구조는 합성곱 신경망 (Convolutional Neural Network, CNN)이다. 이는 이미지 데이터 특성상, 한 부분에 대한 픽셀 분포가 다른 부분들과 동일하다는 가정, 그리고 픽셀 종속성은 특징이 있는 작은 지역에 제한한다는 점이 합성곱 신경망의 학습 방법과 맞기 때문이다[8]. 합성곱 신경망은 AlexNet, GoogleNet, VGG, ResNet, DenseNet, HRNet과 같은 모델들을 거치며 모델 구조의 개선과 더불어 해상도 보존과 같은 기술적 특이점을 연구하며 다양한 컴퓨터 비전 과업에 있어 많이 사용되고 있다[8-13]. 하지만, 최근에는 Attention 기법을 활용한 Transformer 모델이 인공지능 각 분야에서 훌륭한 성과를 보이고 있다. 사람 자세 추정 과업에서는 ViT 구조를 활용한 ViTPose가 자세 추정 벤치마크 MS COCO, OCHuman, MPII에서 최고 성능을 달성했다[15-17]. 그러나 ViT는 구조적으로 Inductive Bias를 고려하기 어렵기에, 매우 큰 Dataset과 많은 파라미터의 수로 이를 극복하고자 하지만, 복잡성이 높아진 모델은 학습과 추론에 있어 더 큰 지연시간을 야기한다. 결과적으로는 환경에 따른 제약이 많아지며, 특히 종단 장치(Edge Device)와 같은 제한된 환경에서의 적용을 어렵게 만든다. 이러한 한계점을 극복하기 위해서는 Transformer기반 모델의 경량화를 연구하는 것이 중요하다 할 수 있다[18].

MobileViT[19]는 합성곱 신경망을 통해 Inductive Bias와 Local Representation을 가지게 되며, 이전된

정보가 Transformer 블록을 통해 사진 전체의 각 부분과 학습하는 모델이다. 위 방법을 통해 ViT계열의 모델보다 적은 수의 매개변수, 간단한 Data 증강 기법으로도 이미지의 특징을 잘 학습하여, 높은 성능을 보여준다. 또한, 위와 같은 특징들이 MobileViT가 종단 장치에서도 좋은 성능을 가져오는 모델로 만들어준다.

본 연구에서는 합성곱 신경망과 ViT의 장점을 결합한 MobileViT를 활용하여 사람의 자세를 추정할 수 있는 모델을 제안한다. 실험결과 ViTPose보다 더 적은 연산량과 매개변수로, 더 가벼운 모델을 선보이며 MS COCO Validation Dataset에서 69.4% AP Score를 달성했다.

이후 논문의 순서는 다음과 같다. 먼저, 2장에서는 관련 연구들을 소개한다. 3장에서는 제안 모델의 구체적인 방법론을 기술하고, 4장에서는 실험 설계와 결과를 통해 제안 방법의 우수성을 서술한다. 5장에서는 결론 및 향후 연구에 관해 설명하며 논문을 마무리한다.

II. Preliminaries

1. ViT for Pose Estimation

ViT를 기반으로 한 ViTPose는 현재 사람 자세 추정의 기준이 되는 Dataset인 MS COCO, OCHuman, MPII에서 최고 성능을 달성했다. ViTPose는 ViT구조를 사용하였고, Masked Image Modeling으로 사전학습 된 초깃값을 통해 좋은 성능을 보여주었다. 그러나 동시에 Transformer는 Inductive Bias가 부족하여 학습에 많은 Dataset과 매개변수가 사용되기 때문에 많은 자원이 필요하다는 단점이 있다. ViT-H 모델의 경우, 약 6억 3천 2백만 개의 매개변수를 사용했으며, 이는 ResNet-50이 약 2천 3백만 매개변수를 사용한 것 대비, 훨씬 큰 규모의 매개변수를 요구하였다. 결과적으로 ViTPose가 자세 추정에서 좋은 성능을 보여주었지만 이를 위해서는 많은 계산 자원이 필요하게 되었다. 본 논문에서는 이와 같은 무겁고 복잡한 구조의 약점을 극복하고자 MobileViT를 사용하여 모델 규모 측면에서의 개선을 시도하였다.

2. MobileViT

MobileViT는 합성곱 신경망과 ViT의 강점을 결합한 모델이다. 합성곱 신경망을 통해 적은 수의 매개변수로 Inductive Bias가 포함된 Local Representation을 학습하고, Transformer 블록을 통해 Global Representation을 학습한다. 위 과정을 통해 많은 Data와 매개변수가 필요하다는 ViT의 한계점을 극복함과 동시에 좋은 성능을 보여주었다. 합성곱 신경망과 ViT기반의 모델 성능을 비교한 결과, DeIT는 약 5백 7십만 개의 매개변수와 1.3GFLOPs로 Top-1 정확도 72.2%를, MobileViT는 약 2백3십만 개의 매개 0.7GFLOPs로 Top-1 정확도 74.8%를 달성했고, 합성곱 신경망을 기반으로 모델 MobileNetV3에서도 3.2% 높은 성능을 보여주었다. 이처럼 가벼우면서 뛰어난 성능을 보인 MobileViT는 현실에서의 다양한 문제, 과업들에 접목되어 그 성능을 발휘하고 있다. 예시로, Object Detection의 세부 분야 중 하나인 Drone Detection에서 사용된 YOLOv4-MCA[20] 모델은 VggNet16을 백본으로 하며, 약 2천만 개의 매개변수를 통해 초당 43개의 프레임을 처리하였다. 하지만, 백본을 MobileViT로 진행했을 때, 약 1천 3백만 개의 매개변수를 사용하여, 초당 40개의 프레임을 처리할 수 있었다. 이는, MobileViT가 Real-Time Detection에서 연산량 대비 성능이 더 효율적인 모습을 보여준다. 다른 경우로는, Image Segmentation을 통해 COVID-19의 양성과 음성을 분류하는 연구에서 MMViT-Seg[21]가 좋은 성능을 보여줬다. 모델은 MobileViT와 CNN 구조의 결합으로 약 1백만 개의 매개변수와 4.6GFlops의 규모를 가졌다. 위 수치는 해당 연구의 비교 모델들 대비 평균적으로 약 1천 5백만 개의 매개변수와 21GFlops가 적었으며, 0,078MAE(Mean Absolute Error)로 가장 좋은 성능을 보였다. 이처럼 MobileViT는 효율적인 구조에서 나오는 적은 연산량과 좋은 성능이 현재 다양한 과업들에서 적용되고 있지만, 최근까지 MobileViT를 사용한 사람 자세 추정 연구는 진행되지 않았기에, 본 논문에서는 이를 집중적으로 연구하였다.

3. Transfer Learning

현재 딥러닝 분야에서는 모델의 성능을 끌어올리고 세부 과업 수행을 위해 많은 양의 Data가 요구되고 있다. 해당 문제를 극복하기 위한 방법으로 전이학습이 등장하였다. 이는 큰 Dataset을 통해 학습한 모델의 가중치를 먼저 불러온 후 해결하고자 하는 과제에 있어 작은 Dataset으로 가중치를 미세 조정하며 높은 정확도에 더 빠르게 도달하는 방법이다. 컴퓨터 비전 분야에서는 대체로 Imagenet-21k Dataset[22]을 사전학습 하여 모델을 먼저 사전 학습시키고 세부 과제에 맞게 재조정을 진행한다. 사람 자세 추정에서도 이와 같은 전이학습은 많이 활용되고 있으며, 특히 Transformer는 학습에 소요되는 시간이 비교적 많이 소모되기 때문에 이를 통해 효율적으로 학습이 가능하다. 본 연구에서 또한, 전이학습을 사용하여 효과적인 성능 개선을 시도하였다.

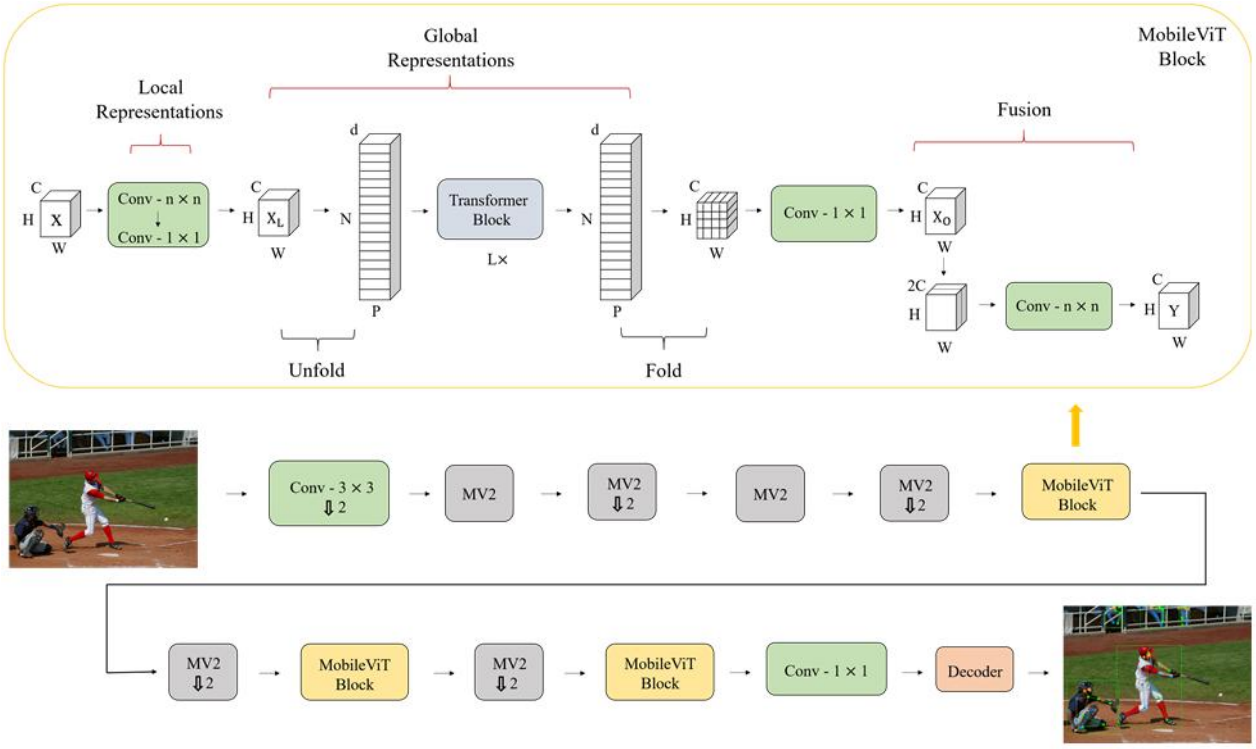

Fig. 1. System Architecture

III. The Proposed Scheme

1. Architecture

본 논문의 목표는 사람 자세 추정에 있어 MobileViT를 사용하여 성능손실을 최소화하고 경량화된 모델 MobileViTPose를 제안하는 것이다. 이를 위해 사용된 모델 MobileViT는 MobileNetV2블록(MV2)과 MobileViT블록의 연속적 조합으로 구성되었으며, Fig. 1과 같다. down-sampling 블록의 경우↓2로 표기하였다.

입력값 X는 높이(H), 너비(W), 채널(C)의 차원을 가지고 있고 N × N 합성곱 연산(Conv)을 진행하여 출력값을 갖는다. 이를 통해 입력 이미지에 대해서 국소 표현을 학습할 수 있다. MV2 블록을 지나면서 채널의 크기가 절반으로 축소되고, Point-Wise 합성곱 연산을 지나게 되면서 d차원으로 사영된다. 위 과정을 식(1)에 나타낸다.

$$X \in R^{H \times W \times C},\ X_L = Conv_{n \times n}(X),$$
$$X_L \in R^{\frac{H}{2} \times \frac{W}{2} \times d} \quad ...(1)$$

식(2)에서는 MobileViT 블록에서 국소 표현된 특징 맵을 이미지와 같은 차원 Patches(X_U)로 출력한다. X_U의 차원 P=WH는 각 Patch의 가로(H≤N), 세로(W≤N) 크기이고 N은 입력값 내에 존재하는 Patches의 수이다.

$$X_U \in R^{P \times N \times d} \quad ...(2)$$

각 Patch($X_G(p)$)는 Transformer 블록을 통해 Inner-Patch와 전체적인 표현을 학습한다. 이로써 MobileViT는 Patch 및 픽셀의 순서를 학습하기 때문에 ViT 특유의 Inductive Bias 문제점을 해소하게 된다.

$$X_G(p) = Transformer(X_U(p)),\ 1 \le p \le P,$$
$$X_G \in R^{P \times N \times d} \quad ...(3)$$

이후 식(3)에서 구한 X_G의 차원을 Fold하여 $X_F \in R^{H \times W \times d}$ 만들고, Point-Wise 합성곱 연산을 통해 입력값 X와 동일한 차원을 출력값 X_O을 만든다. 위 과정을 식(4)가 나타낸다.

$$X_F \in R^{H \times W \times d},\ X_O = Conv_{1 \times 1}(X_F),$$
$$X_O \in R^{H \times W \times C} \quad ...(4)$$

마지막으로, 식(5)에서 입력값 X와 산출값 X_L의 Skip Connection을 위해 N × N 합성곱 연산(Conv)을 수행하며, 최종 출력값 Y를 만든다.

$$Y = Conv_{n \times n}(X + X_L),\ Y \in R^{H \times W \times C} \quad ...(5)$$

본 논문에서는 식(6)과 같은 Decoder를 사용하였다. 4번의 쌍선형 보간법과 ReLU함수를 통해 확대된 특징맵을, 3×3크기의 합성곱 연산의 커널을 통해 히트맵(K)을 구했으며, 키포인트의 수(N_K)만큼 차원을 가지게 된다.

$$K = Conv_{3 \times 3}(Bilinear(ReLU(Y))),$$

$$K \in R^{\frac{H}{4} \times \frac{W}{4} \times N_K} \quad ...(6)$$

2. Scale of the Proposed Model

제안 모델인 MobileViTPose의 구조가 단순하여 Layer와 Feature Dimension을 자유롭게 조절하며 모델의 규모를 쉽게 바꿀 수 있다. 이와 같은 특징을 활용하여 본 연구에서는 560만 개의 매개변수를 가지며 모델 규모가 큰 MobileViT-S와 130만 개의 매개변수를 갖는 작은 사이즈 MobileViT-XXS를 통해 모델의 성능을 확인했다.

IV. Experiments and Results

본 장에서는 제안하는 네트워크의 성능을 종합적으로 판단하기 위해 다음과 같은 Research Question(RQ) 두 가지를 설정 후 실험을 통해 답을 찾는다.

RQ1 : 제안된 MobileViTPose가 최신 사람 자세 추정 모델들을 상대로 한 성능 경쟁력은 충분한가?

RQ2 : 사전 학습을 통한 초기 가중치 갱신의 영향력이 중요한가?

1. Dataset

본 연구는 MS COCO Dataset을 통해 실험이 진행되었다. 해당 Dataset의 통계자료는 Table 1과 같다.

Table 1. Dataset

Dataset	Train	Validation	Test	Classes
COCO	118,287	5,000	40,670	17

MS COCO Dataset은 ImageNet-1K Dataset의 문제점을 해결하기 위해 2014년 제안되었다. 150만 개의 사물 객체와 열일곱 개의 사람 자세를 가지고 있다.

2. Evaluation

본 연구에서의 평가지표로는 Average Precision(AP), Average Recall(AR) 두 가지를 사용하였다. AP, AR은 사람 자세 추정 평가에서, 추정 자세와 정답 자세의 유사성을 나타내는 척도로 사용되며, Object Keypoint Similarity(OKS)에 의해 계산된다. OKS는 각주 된 관절의 추정 좌표와 정답 좌표의 유사성 평균을 나타낸다.

$$OKS = \frac{\sum_i \exp(-\frac{d^2}{2s^2k^{2_i}})\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad ...(7)$$

식(7)에서 d^i는 관절 점 i의 추정 좌표와 정답 좌표의 거리를 의미하고, S는 인물의 크기, k^i는 관절 점의 종류마다 설정되는 상수이며 마지막으로 v^i는 관절 점이 주석 여부를 나타낸다.

제안 모델 MobileViTPose의 우수성을 확인하기 위해 아래 4가지 모델들의 변형들과 성능 비교를 진행하였다. 먼저, ViTPose는 Transformer 모델을 비전 과업에 맞게 설계한 ViT 인코더를 가지고 사람 자세 추정을 수행한 모델이며, 현재 여러 벤치마크에서 최고성능을 달성한 모델이다. 다음으로, HRNet은 컴퓨터 비전 과업에서 중요시되는 고화질 특징맵에 집중하여, 신경망의 층이 늘어남에도 지속적으로 고해상 특징정보를 유지하는 병렬형 네트워크이다. 세 번째, VGG는 3×3 합성곱 필터를 통한 CNN 기반의 깊은 신경망 모델이다. 마지막으로, SimpleBaseline은 ResNet기반의 특징 추출과 확장 구조를 갖는 신경망이다.

3. Parameter Settings

MobileViT 는 사람 자세 추정의 공통적인 Top-Down 방식을 따른다. 본 논문에서는 MobileViT-S, MobileViT-XXS 신경망을 사용하여 MobileViTPose-S, MobileViTPose-XXS을 제안했으며, MMPose Codebase[23] 안에서 NVIDIA Geforce RTX 3060으로 학습하였다. 신경망은 ImageNet-1k Classification Dataset[24]으로 초기화된 사전 학습 가중치를 사용하였다. MobileViTPose를 학습하기 위한 MMPose의 기본적인 실험 환경으로 256×256 입력 해상도와 AdamW Optimizer[25], 5×10^{-4}학습률을 사용하였다. 모델은 210epoch 학습하며 170번째 와 200번째에서 각각 학습률이 10배씩 감소하였다[26].

4. Results

4.1 Experimental Results of RQ1

Table 2는 비교 모델과 제안 모델의 성능을 나타낸다. 비교를 위한 CNN 계열의 모델로는 ResNet의 백본과 Deconvolution Head 네트워크의 결합을 통해 간단한 구조를 만든 SimpleBaseline(ResNet-50), 이미지의 다양한 해상도를 병렬적으로 처리한 HRNet(W32), 합성곱 층과 풀링 층의 반복적인 구조를 통해 깊은 네트워크를 표현한 VGG가 있으며, 제안 모델 MobileViTPose(S)사이에서의 성능 차이는 AP Score 기준, 각각 1%P, 5%P, 0.4%P 낮지만, 비교대상 측 모델들의 매개변수의 개수에서 73%, 68%, 52%, Flops에서는 각각, 66%, 62%, 81%

Table 2. Comparative Experiment Results Table

Model	Backbone	Params (M)	Flops (G)	Input Resolution	Feature Resolution	COCO val	
						AP	AR
VGG	VGG16	19	16	256×192	1/4	69.8	75.4
SimpleBaseline	ResNet-50	34	9	256×192	1/4	70.4	76.3
SimpleBaseline	ResNet-152	69	16	256×192	1/4	72.0	77.8
HRNet	HRNet-W32	29	8	256×192	1/4	74.4	78.9
HRNet	HRNet-W32	29	8	384×288	1/4	75.8	81.0
HRNet	HRNet-W48	64	16	256×192	1/4	75.1	80.4
HRNet	HRNet-W48	64	16	384×288	1/4	76.3	81.2
ViTPose-B	ViT-B	86	17	256×192	1/16	75.8	81.1
*MobileViTPose-XXS	MobileViT-XXS	4	2	256×256	1/4	61.7	67.9
*MobileViTPose-S	MobileViT-S	9	3	256×256	1/4	69.4	74.9

Table 3. Comparative Experiment Results Table

Model	AP	AP_{50}	AP_{75}	AR	AR_{50}	AR_{75}
MobileViTPose-XXS(Plain)	60.2	84.8	66.7	66.5	89.4	72.7
*MobileViTPose-XXS(Pre-trained)	61.7	86.1	69.0	67.9	90.5	74.6
MobileViTPose-S(Plain)	63.9	86.2	70.8	69.9	90.4	76.3
*MobileViTPose-S(Pre-trained)	69.4	88.8	77.0	74.9	92.8	82.0

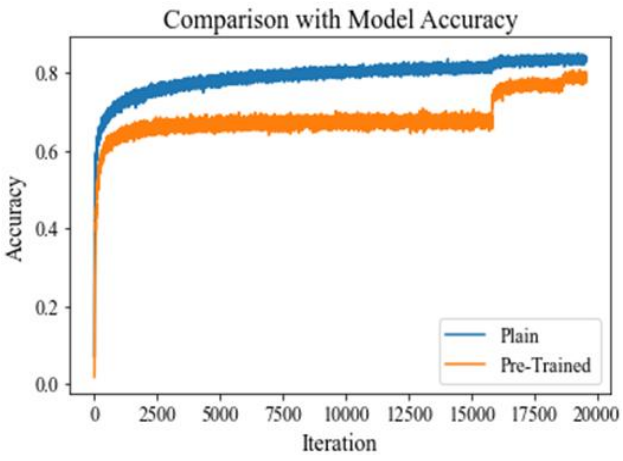

Fig. 2. Comparison with Model Accuracy

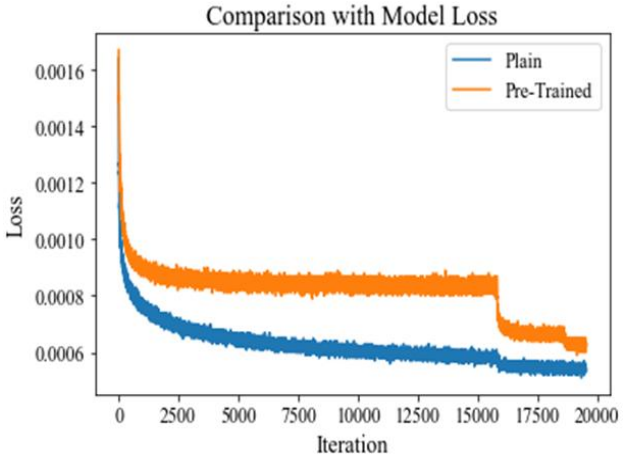

Fig. 3. Comparison with Model Loss

더 적은 수치를 보여주었다. 한편, 특징 학습에 있어 ViT 네트워크를 백본으로 가지며, 자세 추정을 위한 Deconvolution Head 네트워크를 결합한 모델인 ViTPose(ViT-B)는 제안 모델 MobileViTPose(S)와의 성능 차이에서, AP Score가 6.4%P 낮지만, ViTPose-B의 매개변수에서 90%, Flops에서는 83% 더 적은 수치를 보여주었다. 또한, 제안 모델의 규모 변화 실험을 통해 MobileViTPose-XXS에서 S로의 모델 크기가 증가함에 따라 MobileViTPose의 성능이 향상되는 것을 보아, MobileViTPose의 우수한 확장성과 유연성을 확인할 수 있었다.

4.2 Experimental Results of RQ2

Table 3, Fig. 2, Fig. 3을 통해, COCO Dataset으로 바닥부터 학습시킨 기본 모델과, ImageNet-1K로 사전학습 된 초기 가중치 값을 COCO Dataset에 대해 적응시킨 모델의 결과를 비교할 수 있다. ImageNet-1K Dataset을 가지고 분류 작업을 수행한 사전 학습된 가중치는 기본 모델보다 적응 학습에서 이미지의 색, 선과 같은 일반적인 특징 정보를 더 잘 반영하였다고 해석할 수 있다. Fig. 2를 통해 기본 모델보다 초기에 훨씬 높은 정확도와, 낮은 손실률을 기록하는 걸 확인할 수 있으며, 학습이 완료된 210epoch의 결과에서도 5.5%P 높은 AP Score를 달성했다. 이하 Fig. 4를 통해 제안 모델 MobileViTPose를 통한 MS COCO Dataset에서의 실험

Fig. 4. Mobile ViTPose Experimental Results Photos

결과 사진들을 수록한다. 사진 속 사람 자세 추정 좌표들은 인물이 여럿이거나 역동적인 모습, 인물 크기, 사진의 종횡비와 상관없이 관절 좌푯값을 잘 포착하고 있다.

V. Conclusions

본 논문에서는 효율적이면서도 높은 성능을 보이는 MobileViTPose를 제안하였다. MobileViTPose는 MobileNetV2 Block과 MobileViT Block의 연속적인 조합으로 구성된 하이브리드 백본 구조와 사람 자세 추정을 위한 Deconvolution Head 네트워크를 결합하였다. 제안 모델 중 더 큰 사이즈인 MobileViTPose-S는 3.28GFLOPs와 972만 매개변수를, 작은 사이즈인 MobileViTPose-XXS는 1.78GFLOPs와 442만 매개변수를 가졌다. MS COCO validation set에서의 실험 결과는 각각 69.4, 61.7 AP Score를 달성하였다. 또한, 사전학습에 ImageNet-1k Classification Dataset으로 초기화된 가중치를 사용했을 경우, 기본 모델들보다 5.5%P, 1.5%P 높은 AP Score와 1×10^{-4} 더 낮은 손실률을 기록하였다. 향후 연구에서는 다양한 이미지 해상도에 대해 (256×192, 384×288) 성능을 확인할 것이며, 사람 자세 추정을 위한 신경망 구조를 개량할 것이다.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2022-00166634).

REFERENCES

[1] B. Xiao, H. Wu, and Y. Wei, “Simple baselines for human pose estimation and tracking,” Proceedings of the European

Conference on Computer Vision, pp. 466-481, Apr. 2018. DOI: 10.48550/arXiv.1804.06208

[2] J. Park, D. Park, D. Hwan, Y. Na, S. Lee, “Deep-Learning Based Real-time Fire Detection Using Object Tracking Algorithm,” Proceedings of The Korea Society of Computer and Information, Vol. 27, No. 1, pp. 1-8, Jan. 2022. DOI:10.9708/jksci.2022.27.01.001

[3] D. Hwang, G. Moon, Y. Kim, “SKU-Net: Improved U-Net using Selective Kernel Convolution for Retinal Vessel Segmentation,” Proceedings of The Korea Society of Computer and Information, Vol. 26, No. 4, pp. 29-37, Apr. 2021. DOI:10.9708/jksci.2021.26.04.029

[4] S. Yang, S. Lee, “Improved CNN Algorithm for Object Detection in Large Images,” Proceedings of The Korea Society of Computer and Information, Vol. 25, No. 1, pp. 45-53, Jan. 2020. DOI:10.9708/jksci.2020.25.01.045

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, Aidan, N. Gomez, Ł. Kaiser, I. Polosukhin, “Attention Is All You Need,” Proceedings of the Neural Information Processing Systems, Dec. 2017. DOI: 10.48550/arXiv.1706.03762

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. “An Image Is Worth 16X16 Words: Transformers For Image Recognition At Scale,” Proceedings of the International Conference on Learning Representations, Aug. 2021 DOI: 10.48550/arXiv.2010.11929

[7] H. Touvron, M. Cord, M. Douze. F. Massa, A. Sablayrolles, H. J´egou, “Training data-efficient image transformers & distillation through attention,” arXiv preprint arXiv:2012.12877, Dec. 2022. DOI: 10.48550/arXiv.2012.12877

[8] A.. Krizhevsky, I. Sutskever, Geoffrey E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” Proceedings of the Neural Information Processing Systems pp. 84-90, 2012. DOI: 10.1145/3065386

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, “Going Deeper with Convolutions,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June. 2015. DOI: 10.1109/CVPR.2015.7298594

[10] K. Simonyan, A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” arXiv preprint arXiv:1409.1556, Apr. 2014. DOI: 10.48550/arXiv.1409.1556

[11] K. He, X. Zhang, S. Ren, J. Sun, “Deep Residual Learning for Image Recognition,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778. June. 2016. DOI: 10.1109/CVPR.2016.90

[12] G. Huang, Z. Liu, L. van der Maaten, Kilian Q. Weinberger, “Densely Connected Convolutional Networks,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700-4708, July. 2017. DOI: 10.1109/CVPR.2017.243

[13] J. Wang, Ke Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, B. Xiao, “ Deep High-Resolution Representation Learning for Visual Recognition,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3349-3364, Apr. 2021. DOI: 10.1109/TPAMI.2020.2983686

[14] Y. Xu1, J. Zhang, Q. Zhang, D. Tao. “ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation,” Proceedings of the Neural Information Processing Systems, Oct. 2022. DOI: 10.48550/arXiv.2204.12484

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” Proceedings of the European Conference on Computer Vision, May. 2014. DOI: 10.48550/arXiv.1405.0312

[16] S.-H. Zhang, R. Li, X. Dong, P. Rosin, Z. Cai, X. Han, D. Yang, H. Huang, and S.-M. Hu. “Pose2seg: Detection free human instance segmentation,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 889−898. June. 2019. DOI: 10.1109/CVPR.2019.00098

[17] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “2d human pose estimation: New benchmark and state of the art analysis,” Proceedings of the Conference on Computer Vision and Pattern Recognition, pp. 3686−3693. June. 2014. DOI: 10.1109/CVPR.2014.471

[18] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollar, and R. Girshick, “Early convolutions help transformers see better,” arXiv preprint arXiv:2106.14881. June. 2021. DOI: 10.48550/arXiv.2106.14881

[19] D. Mehta, M. Rastegari, “Mobilevit: Light-Weight, General-Purpose, And Mobile-Freindly Vision Transformer,” Proceedings of the International Conference on Learning Representations. Jan. 2022. DOI: 10.48550/arXiv.2110.02178

[20] Q. Cheng, X. Li, B. Zhu,Y. Shi, B. Xie, “Drone Detection Method Based on MobileViT and CA-PANet,” Proceedings of the Electronics, pp. 223-239. Dec. 2023. DOI: 10.3390/electronics12010223

[21] Y. Yang, L. Zhang, L. Ren, X. Wang, “MMViT-Seg: A lightweight transformer and CNN fusion network for COVID-19 segmentation,” Proceedings of the Computer Methods and Programs in Biomedicine, Mar. 2023. DOI: 10.1016/j.cmpb.2023.107348

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” Proceedings of the Conference on Computer Vision and Pattern Recognition, pp. 248−255, June. 2009. DOI: 10.1109/CVPR.2009.5206848

[23] M. Contributors, “Openmmlab pose estimation toolbox and benchmark,” https://github.com/open-mmlab/mmpose.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, Alexander, C. Berg. Li Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” arXiv preprint arXiv:1409.0575, Sep. 2014. DOI: 10.48550/arXiv.1409.0575

[25] S. J. Reddi, S. Kale, and S. Kumar, “On the convergence of adam and beyond,” Proceedings of the International Conference on Learning Representations, Sep. 2019. DOI: 10.48550/arXiv.1904.09237

[26] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” Proceedings of the Neural Information Processing Systems, June. 2019. DOI: 10.48550/arXiv.1906.08237

Authors

Kunwoo Kim is currently enrolled in AI, Big Data & Management from Kookmin University, Korea. He is interested in deep learning, computer vision, recommendation system, and model weight reduction.

Jonghyun Hong is currently enrolled in AI, Big Data & Management from Kookmin University, Korea. He is interested in deep learning, computer vision, and natural language processing.

Jonghyuk Park received the B.S. and Ph.D. degrees in Industrial Engineering from Seoul University, Korea, in 2015 and 2021, respectively. In 2015, Dr. Park was a Product Management Engineer with Samsung Electronics. From 2020 to 2021, he was a Research Engineer in ai.m, Seoul, Korea. Since 2021, he has been an Assistant Professor at the Department of AI, Big Data & Management, Kookmin University, Seoul, Korea. His current research interests include computer vision, and machine learning applications.