신체부위의 파악이 가능한 변형된 트랜스포머

강성재*, 김성수**, 서기성*
*서경대학교 전자컴퓨터공학부
**연암공대 전기전자공학과

Body-Part-Aware Modified Vision Transformer

Sungjae Kang*, Seong Soo Kim**, Kisung Seo*
*Division of Electronic and Computer Engineering. Seokyeon University
**Dept. of Electrical & Electronic Engineering, Yonam Institute of Technology

Abstract - 최근 딥러닝 네트워크 기반의 사람 재인식 성능을 높이기 위해 Self-Attention 기반의 트랜스포머 네트워크의 사용이 시도되고 있다. 또한 더욱 효과적인 사람의 재인식을 위해 키포인트 추정 모델이 결합된 방식이 제안되고 있다. 하지만 이러한 방식은 학습 때와 마찬가지로 추론 시 키포인트 추정 모델이 필요하다는 단점이 존재한다. 본 논문에서는 추론 시 트랜스포머 모델만을 사용하여 사람 신체 부위를 판별하고, 특히 가려진 사람의 재인식 성능의 향상을 위해 변형된 트랜스포머 구조를 제안한다. 제안 기법의 효과를 검증하기 위해 Occluded-Duke 데이터셋을 사용해 최신의 기법들과 성능을 비교 한다.

1. 서 론

딥러닝 네트워크 기반의 사람 재인식은 다중 감시 시스템에서 특정 인물을 찾아내거나, 자율주행 환경에서의 보행자를 인식하는 등 다양하게 적용되고 있다. 딥러닝 네트워크의 발전에 따라, 최근 가려진 사람의 재식별 문제와 관련된 연구가 활발히 진행되고 있다. 가려지지 않은 신체 부위에 집중하는 연구[1,2], 고차원 그래프를 이용하는 연구[3], 다양한 데이터 증강을 수행하는 연구[4,5]가 수행되고 있다. 한편, 최근 ViT(Vision Transformer)[6]가 소개됨에 따라 트랜스포머 기반의 더욱 발전된 연구[7-9]가 수행되고 있다. 본 논문에서는 키포인트 추정 모델을 기반으로 ViT 모델을 구조적으로 변형하여 신체부위를 추정함과 동시에 가려진 사람의 재인식을 효과적으로 수행할 수 있는 기법을 제안한다.

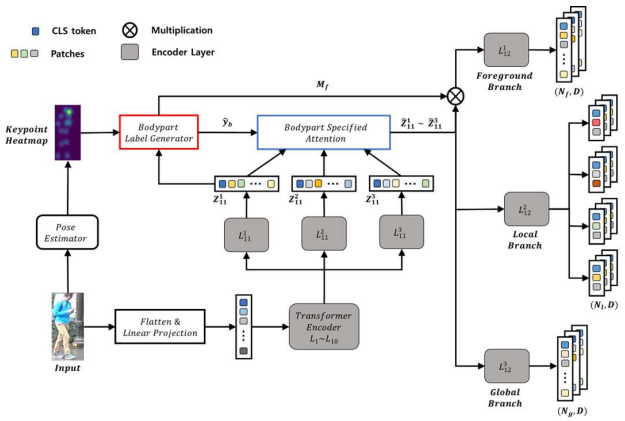
2. 신체부위 인식 가능한 변형된 트랜스포머

2.1 TransReID

TransReID[7]는 ViT[6] 기반으로 사람의 재인식 문제를 효과적으로 해결한 최초의 연구이다. ViT의 입력 구조와 동일하게 입력 이미지를 다수의 패치로 분할하고 각 패치를 저차원의 벡터로 임베딩한다. 임베딩된 패치들을 다수의 트랜스포머 인코더 층을 통과시켜 사람 부분에 집중하도록 한다. TransReID는 ViT와는 다르게 마지막 인코더 층을 2개의 가지로 구성한다. 하나는 전역(global) 가지로서,이전 층에서의 출력 벡터를 모두 입력으로 사용하여 예측을 수행한다. 다른 하나는 JPM(Jigsaw Patch Module)으로, 이전 층에서의 출력 벡터들을 4 개의 그룹으로 분할하여 독립적인 입력으로 사용함으로써, 사람의 부분적인 정보들을 사용하여 예측을 수행한다.

2.2 신체부위 인식 가능한 변형된 트랜스포머

제안 기법의 전체 구성도가 그림 1에 나타나 있다. TransReID[7] 모델에서 네 가지의 구조적 변경이 수행되었다. 첫째, 마지막에서 두 번째 층의 인코더를 3개로 구성하여 동일 인물에 대한 다양한 특징을 추출한다. 둘째, 키포인트 히트맵을 이용하여 신체부위 라벨 생성기에서 각 패치에 해당하는 신체부위를 인식할 수 있도록 한다. 셋째, 마지막에서 두 번째 층에서 추출한 다양한 특징에 각 신체부위의 고유한 특징 값을 주입한다. 넷째, 예측을 수행하기 위해 세 개의 가지를 구성한다.



<그림 1> 전체 구조도

2.3 신체부위 라벨 생성기

추론 시 키포인트 추정 모델 없이 신체부위를 추정할 수 있는 신체부위 라벨 생성기를 제안한다. 이 모듈의 학습은 2단계로 수행된다. 첫 단계로, 키포인트 추정 모델을 통해 얻은 키포인트 히트맵을 전처리 과정을 통해 의사 라벨(Pseudo label)을 생성한다. 두 번째 단계로, 트랜스포머 인코더층의 출력 패치를 추가적으로 구성한 분류기에 입력으로 사용하여 의사 라벨과의 손실함수를 계산한다. 이에 대한 손실 함수는 다음과 같다.

$$L = -\frac{1}{m}\sum_{i=1}^{14} t_i \log(y_i) \quad (1)$$

여기서, t_i는 의사 라벨, y_i는 분류기의 출력값에 소프트맥스 함수를 적용한 값을 나타낸다. 학습이 완료되면 분류기는 키포인트 추정 모델과 동일한 수준의 신체부위 라벨을 생성할 수 있게 된다.

3. 실험 및 결과

3.1. 실험 환경

제안 기법을 검증하기 위해 대표적인 벤치마크 데이터 셋 중의 하나인 Occluded-Duke를 사용한다. Occluded-Duke은 train, query, gallery set으로 나뉘어져 있고, 8대의 카메라를 통해 추출되었다. train set은 702개의 사람 id, 15,618 장의 이미지가 존재한다. query set은 519개의 사람 id, 2,210 장의 이미지가 존재한다. gallery set은 1,110,개의 사람 id, 17,771 장의 이미지가 존재한다.

3.2. 실험 결과

최신의 기법들과 비교한 성능 결과가 표 1에 나와 있다. 제안하는 기법이 기존의 합성곱 신경망 기반의 기법들[1,2]에 비해 상당한 성능 개선이 있음을 볼 수 있다. 또한 트랜스포머 기반의 기법들과[7,9] 비교해서도 우위에 있음을 알 수 있다.

<표 1> Occluded-Duke 데이터에 대한 재인식 실험 결과

Method	mAP (%)	Rank-1 (%)
PVPM[1]	37.7	47.0
HOREID[2]	43.8	55.1
PAT[9]	53.6	64.5
TransReID[7]	55.7	64.2
Ours	59.8	69.0

[참 고 문 헌]

[1] S. Gao, J. Wang, H. Lu, Z. Liu, “Pose-guided visible part matching for occluded person reid,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11744 - 11752, 2020.

[2] J. Yang, J. Zhang, F. Yu, X. Jiang, M. Zhang, X. Sun, Y.C. Chen, W.S. Zheng, “Learning to know where to see: A visibility-aware approach for occluded person re-identification,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11885 - 11894, 2021.

[3] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, J. Sun, “High-order information matters: Learning relation and topology for occluded person re-identification,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6449 - 6458, 2020.

[4] P. Chen, W. Liu, P. Dai, J. Liu, Q. Ye, M. Xu, Q. Chen, R. Ji, “Occlude them all: Occlusion-aware attention network for occluded person re-id,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11833 - 11842, 2021.

[5] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, “Random erasing data augmentation,” in Proceedings of the AAAI conference on artificial intelligence,” vol. 34, pp. 13001 - 13008, 2020.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, “An image is worth 16x16 words: Transformers for image recognition at scale,” in International Conference on Learning Representations, 2020.

[7] S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, “Transreid: Transformer based object re-identification,” in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15013 - 15022, 2021.

[8] M. Jia, X. Cheng, S. Lu, J. Zhang, “Learning disentangled representation implicitly via transformer for occluded person re-identification,” in IEEE Transactions on Multimedia, 2022.

[9] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, F. Wu, “Diverse part discovery: Occluded person re-identification with part-aware transformer,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.2898 - 2907, 2021.