https://doi.org/10.7840/kics.2023.48.9.1175

Transformer 기반 인물 재식별을 위한 상대적 위치 임베딩 활용방법

김 성 수•, 김 경 환°

A Relative Positional Embedding Scheme for Transformer-Based Person Re-Identification

Seong-Su Kim•, Gyeonghwan Kim°

요 약

본 논문은 상대 위치 임베딩을 활용한 transformer 기반의 인물 재인식 모델의 학습 방법을 제안한다. 이미지의 시각적 정보에 의존하는 기존 방식의 한계점을 보완하기 위해, 인물의 신체 구조의 형태적 특성을 상대 위치 임베딩을 통해 정의하고, 이들을 추가적인 근거로 활용한다. 제안한 방식의 적용으로 기존 방식 대비 5개의 인물 재인식 datasets에 대해 정량적, 정성적으로 성능이 향상됨을 확인 하였다.

Key Words : Person re-identification, Transformer, relative positional embedding

ABSTRACT

In this letter, we propose a training scheme for a transformer-based person re-identification model using relative positional embeddings. To overcome the limitations of existing methods that rely on the visual information of an image, we define the topological and positional characteristics of a person's body structure through relative positional embeddings and uses them as an additional cue. In a set of experiment conducted for five popular person ReID benchmark datasets, the proposed scheme brings promising improvement.

Ⅰ. 서 론

인물 재인식(Person Re-Identification)은 지능형 다중 CCTV 시스템의 서로다른 카메라에 등장하는 같은 인물을 찾는 것을 목표로 한다. 같은 인물이라 하더라도 카메라 시점에 따른 인상착의의 변화가 발생 하거나 신체 일부가 가려지는 등의 제약 상황이 발생하기 때문에 인물의 세밀한 특징을 파악하는 것이 중요하다.

최근 vision transformer[1] 기반의 인물 재인식 연구 중에서 TransReID-SSL[2]과 Valid key point[3]는 이미지를 여러 개의 작은 영역으로 구분한 후, 각 영역에 포함된 패치에서 특징을 뽑아 통합하여 인물의 특성을 파악한다.

본 논문에서는 이러한 transformer 기반 모델의 구조적 특징을 활용하여, 인물의 신체 구조를 반영하는 재인식 모델의 학습 방식을 그림 1과 같이 제안한다. 신체의 대표적인 특징을 포함하는 이미지 패치들을 선별 후, 그들 간의 상대적 위치 관계를 relative positional embedding (RPE[4])을 통해 학습한 후, 학습된 RPE를 분포 행렬의 형태로 통합 하여 같은 인물에 대해 일관된 분포가 형성될 수 있도록 손실 함수를 적용한다.

Ⅱ. 제안하는 방법

객체의 구조적 특성은 패치들 간의 상대적 위치 관례를 통해 나타낼 수 있다. 특히, 다양한 종류의 객체를 포함하는 일반적인 이미지 dataset과 달리 인물 재인식 dataset은 사람의 이미지만을 다룬다는 점에서, dataset 전반에 객체의 일관된 구조적 특성이 존재한다. 이는 인물에 상관없이 인체의 구조적 특성이 동일하게 유지되기 때문이다. 다음은 제안하는 방법의 각 단계를 설명한다.

인물의 대표 특징 패치 선택: 그림 1의 A 영역은 인물의 신체적 특성을 대표하는 패치를 추출하는 과정을 보여준다. 각 출력 패치 토큰과 클래스 매개변수간 내적을 계산한 후, 유사도가 높은 패치들을 선별한다.

※ 본 연구는 행정안전부/국토교통과학기술진흥원의 지원으로 수행되었음(과제번호 22PQWO-C153369-04).
• First Author : (ORCID:0009-0000-4480-8560) Department of Electronic Engineering, Sogang University, pand_a@naver.com, 학생(석사과정), 정회원
° Corresponding Author : (ORCOD:0000-0002-7295-2006) Department of Electronic Engineering, Sogang University, gkim@sogang.ac.kr, 정교수, 정회원
논문번호 : 202304-068-C-LU, Received April 4, 2023; Revised May 10, 2023; Accepted May 10, 2023

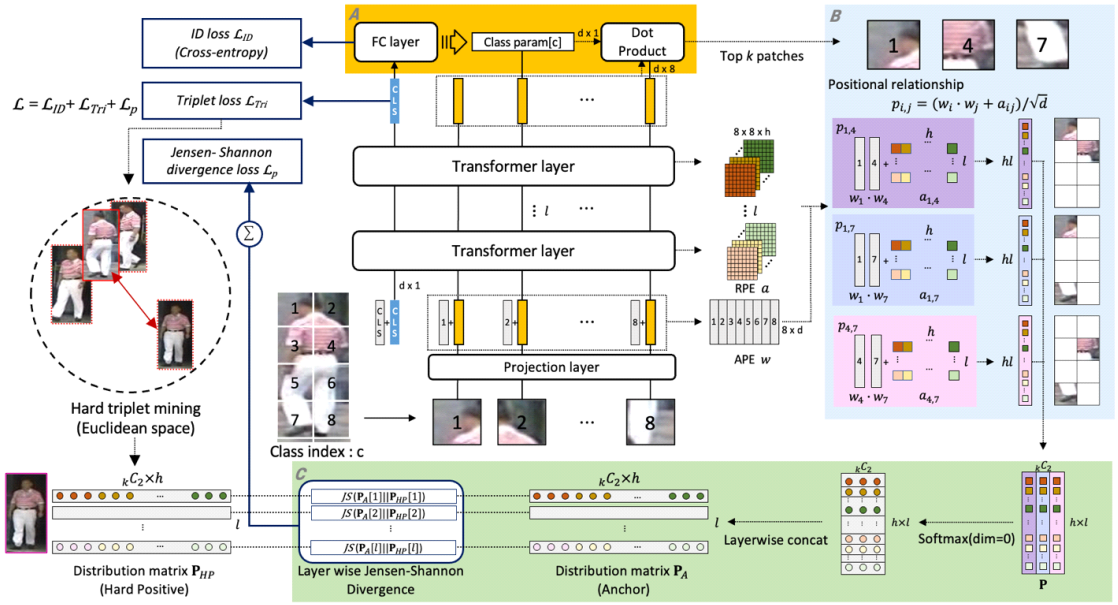

그림 1. 제안하는 인물 재인식 모델의 학습 파이프라인: Transformer 각 계층에 적용한 RPE와 출력 이미지 패치를 활용한다. 방법의 직관적인 설명을 위해 4 × 2 개의 패치로 나눈 경우를 묘사하였다.
Fig. 1. The overall training pipeline of the proposed model - utilization of the relative positional embeddings in each transformer layer and the output image patches in the last layer.

클래스 매개변수는 분류기의 마지막 계층을 구성하며, 인물 이미지가 어떤 클래스에 속하는지 확률을 출력하는데 사용된다. 따라서, 클래스 파라미터는 인물의 대표적인 특징을 담고 있으며, 이와 유사도가 높은 패치일수록 주요 신체 특성을 담고 있을 확률이 높다.

RPE를 통한 신체 구조적 특성 설계: 그림 1의 B영역에 묘사된 것과 같이 선택된 패치들 사이의 상대적인 위치 관계를 RPE로부터 추출하고 이를 통합하여 신체 구조적 특성을 모델링한다. 식(1)은 각 multi-attention head에 적용되는 self-attention 메커니즘에 RPE를 도입할 때의 패치 쌍 (i,j) 간의 attention score e_{ij}를 나타낸다.

$$e_{ij} = softmax\left(\frac{((t_i + w_i)W^Q)\left((t_j + w_j)W^K\right)^T + a_{ij}}{\sqrt{d}}\right) \quad (1)$$

여기서 t = semantic feature, w = absolute positional embedding, d = feature dimension, W^Q, W^K = query, key 변환 행렬, $a_{ij} = (i,j)$ 패치 쌍 간의 RPE를 나타낸다. RPE는 두 패치의 query, key간 내적에 bias의 형태로 더해진다. 이는 각 layer의 모든 head에 대해 독립적으로 존재한다. 식(1)에서 패치 쌍간 위치 관계를 담은 정보는 $p_{ij} = (w_i \cdot w_j + a_{ij})/\sqrt{d}$로 추출하여 표현할 수 있으며, k개의 대표 패치를 선별했을 때, ${}_kC_2$개의 패치쌍이 존재하게 된다. 이 패치쌍의 위치 관계 정보를 통합한 후 정규화 한 분포행렬 $P = softmax\left(concat\ (p_{12}, p_{13}, \ldots, p_{k-1k})\right)$는 각 신체 부위를 대표하는 모든 패치들 간의 위치 관계를 포함 하며 인물의 신체 구조적 특성을 나타낸다.

신체 구조 일관성 학습: 인물의 신체 구조는 각 인물의 고유한 특성이므로, 이를 학습하는 것은 동일 인물을 판단하는데 중요한 근거로 사용될 수 있다. 따라서, 같은 인물의 이미지라면 분포 행렬 P 의 형태가 일관되게 유지될 수 있도록 Jensen-Shannon divergence loss를 적용한다. 그림 1의 C 영역은 이 과정을 나타낸다. Jensen-Shannon divergence loss는 식(2)와 같이 정의되며 두 확률 분포 p, q가 주어졌을 때 두 분포가 유사할수록 값이 작다.

$$JS(p||q) = \frac{1}{2}KL(p||\frac{p+q}{2}) + \frac{1}{2}KL(q||\frac{p+q}{2}) \quad (2)$$

따라서, 동일 인물의 앵커(anchor)와 하드-포지티브(hard-positive) 샘플에 대해 생성된 분포 P간 손실 함수 $L_p = \sum_{i=1}^{l} JS(P_A[i]||P_{HP}[i])$를 최소화한다. 최종 손실 함수는 $L = L_{ID} + L_{Tri} + L_p$로 정의하는데, 여기서

L_{ID}와 L_{Tri}는 각각 엔트로피 손실과 삼중항 손실을 의미하고, l = layer 수, i = layer index, P_A = 앵커의 분포 행렬, P_{HP} = 하드-포지티브의 분포 행렬 이다.

Ⅲ. 실험 및 고찰

표 1에 제시된 인물 재인식 dataset을 사용하여 실험을 수행하였다. 그 중 MSMT17-V2[5]와 Occluded-Duke[6]는 신체 일부가 가려지거나 모자이크 처리된 이미지를 가지고 있어 제안한 방법의 견고성을 검증하기 위해 추가했다. 기준 모델인 TransReID-SSL[2]과의 성능 비교를 통해 제안한 방법의 효용성을 검증하였다.

실험 과정에서 입력 이미지는 384 × 128의 크기로 조정되며 패치의 크기는 16 × 16으로 고정된다. 배치의 크기는 16명의 인물 당 4개의 이미지를 사용하여 64로, 선택하는 패치의 수(k)는 5로 설정한다.

인물 재인식 모델의 성능을 평가하기 위해 R1 score와 mAP를 사용하였다. 기준 모델의 성능은 기준 모델의 논문에서 인용하거나 기준 모델의 공식코드를 재현하여 측정하였다.

그림 2는 제안한 방법이 기준 모델보다 개선된 결과를 제공하는 쿼리 이미지에 대해 rank 1 이미지 쌍과 attention map을 보여준다. 빨간 테두리의 이미지는 기존 방식[2]에서 다른 인물이 매칭된 경우를 나타내며, 초록색 테두리의 이미지는 제안한 방식을 적용 후 동일 인물을 인식한 결과이다. 제시된 예시의 attention 영역을 보면 기준 모델에 의한 결과는 모델의 옷의 색상과 같은 시각적 힌트에 의존하기 때문에 위치관계가 크게 고려되지 않는 반면, 제안된 방식에 의한 결과는 더 정확한 attention 영역을 구성할 뿐만 아니라 실선으로 표시된 것처럼 상대적인 위치 관계를 보존한다.

표 1에서 알 수 있듯이 본 논문에서 제안한 방식을 사용한 경우 모든 dataset에 대해 우수한 성능을 보였으며, 특히 특수한 제약 조건이 존재하는 MSMT17-V2[5]와 Occluded-Duke[6] dataset에서 더 높은 성능 향상을 보인다. 이 결과에 따르면 제안된 방법이 다양한 제약 상황 속에서도 동일 인물을 재인식하는데 효과적임을 확인할 수 있다.

표 1. 기존 방식[2]와 제안한 방식 간 성능 비교
Table 1. The performance comparison between the baseline[2]

Datasets	Performance			
	Baseline		Proposed method	
	mAP	Rank1	mAP	Rank1
Market1501[7]	93.2[2]	96.7[2]	93.45	97.06
DukeMTMC-ReID [8]	84.25	92.32	84.54	92.37
CHUK03-np[9]	88.42	89.9	88.9	89.9
Occluded-Duke[6]	62.75	72.44	63.5	72.94
MSMT17-V2[5]	73.47	88.09	74.1	88.7

그림 2. (a) 기존 방식. (b) 제안한 방식. 기존 방식[2]보다 개선된 결과를 보이는 query-rank 1 이미지 쌍에 대한 attention map 시각화. 검은 선은 각 이미지 쌍 매칭 시 대응되는 attention 영역을 표시.
Fig. 2. Attention maps of pairs of query and rank 1 images for cases where the proposed scheme improved results against the baseline[2], for the same query images. Solid black lines refer to corresponding attentive regions in each image pair.

Ⅳ. 결 론

본 논문은 transformer 기반의 인물 재인식 모델에서 RPE[4]를 활용하여 인물의 신체 구조를 학습하는 방법을 제안하였다. 인물의 주요 신체 부위 간 상대적 위치 정보를 인물을 구별하는 추가적인 단서로 활용하여, 5개의 재인식 datasets에 대해 우수한 성능을 보인다. RPE가 이미지에서 객체의 구조적 특성을 학습하는 데에 활용될 수 있다는 점에서, 제안하는 방법은 인물 외에 일관된 형태를 가진 객체를 인식하고 구분하는 데에 효과적일 것으로 판단된다.

References

[1] A. Dosovitskiy, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. (https://doi.org/10.48550/arXiv.2010.11929)

[2] H. Luo, et al., "Self-supervised pre-training for transformer-based person re-identification," *arXiv preprint arXiv:2111.12084*, 2021. (https://doi.org/10.48550/arXiv.2111.12084)

[3] S. Kim, et al., "Valid keypoint augmentation based occluded person re-identification," *Trans. KIEE*, vol. 71, no. 7, pp. 1002-1007, 2022. (https://doi.org/10.5370/KIEE.2022.71.7.1002)

[4] P. Shaw, et al., "Self-attention with relative position representations," in *Proc. NAACL-HLT*, pp. 464-468, 2018. (https://doi.org/10.18653/v1/N18-2074)

[5] L. Wei, et al., "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE Conf. CVPR*, pp. 79-88, 2018. (https://doi.org/10.1109/CVPR.2018.00016)

[6] J. Miao, et al., "Pose-guided feature alignment for occluded person re-identification," in *Proc. IEEE/CVF ICCV*, pp. 542-551, 2019. (https://doi.org/10.1109/ICCV.2019.00063)

[7] L. Zheng, et al., "Scalable person re-identification: A benchmark," in *Proc. IEEE ICCV*, pp. 1116-1124, 2015. (https://doi.org/10.1109/ICCV.2015.133)

[8] Z. Zheng, et al., "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE ICCV*, pp. 3754-3762, 2017. (https://doi.org/10.1109/ICCV.2017.405)

[9] W. Li et al., "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. CVPR*, pp. 152-159, 2014. (https://doi.org/10.1109/CVPR.2014.27)