사전학습된 CNN의 계층별 특징추출능력연구

Layer-wise Feature Extraction Capacity using Pre-trained CNN

이 재 환[1], 윤 숙[2], 박 동 선[3*]
전북대학교 전자공학과[1,3],
목포대학교 실감미디어기술연구소[2],
목포대학교 멀티미디어 공학과[2],
전북대학교 IT융복합연구센터[3]

Jaehwan Lee[1], Sook Yoon[2], Dong Sun Park[3*]
Department of Electronic Engineering, Chonbuk National University, Korea[1,3]
Research Institute of Realistic Media and Technology, Mokpo National University[2]
Department of Multimedia Engineering, Mokpo National University, Korea[2]
IT Convergence Research Center, Chonbuk National University, Korea[3]

요약

최근 객체인식 분야에서는 Convolutional Neural Network (CNN)이 주목받고 있다. CNN의 특징 중 하나는 입력이미지로 부터 특징 추출 방법을 스스로 학습한다는 것이다. 전통적은 객체인식 방법에서는 hand-written feature extractor를 사용하지만, CNN은 스스로가 특징을 추출한다. 하지만 CNN은 많은 학습데이터와 학습 시간을 필요로 한다. 우리는 객체인식 데이터로 사전학습된 CNN을 사용하여 특징을 추출하였고, 이 특징으로 People re-identification을 수행하였다. 이 과정에서 어떠한 학습도 하지 않았지만 CNN은 다른 영상처리 응용에 대해서도 비교적 좋은 성능을 보여주었다.

Ⅰ. 서론

People re-identification(People re-id)은 두 영상의 사람이 같은 사람인지 판단하는 것이다. 일반적으로 지능형 감시 시스템에서 두 영상은 서로 중첩되지 않은 카메라에서 획득한 영상이다. People re-id에서는 빛의 변화, 촬영 방향의 변화, 카메라 파라미터의 변화, 포즈의 변화, 겹침 현상(Occlusion), 낮은 해상도 등 많은 문제를 겪고 있다.

최근 객체인식 분야에서는 Convolutional Neural Network (CNN)이 주목받고 있다. CNN의 특징 중 하나는 입력이미지로 부터 특징 추출 방법을 스스로 학습한다는 것이다. 전통적은 객체인식 방법에서는 hand-written feature extractor를 사용하지만, CNN은 스스로가 특징을 추출한다.

CNN은 많은 학습 데이터를 필요로 하지만 People re-id 데이터 셋의 학습 데이터는 그리 많지 않다. 이 논문에서 우리는 학습시킨 CNN의 특징 추출 능력이 다른 응용에 대해서도 효과가 있는지 실험하였다. 객체인식 데이터로 pre-train된 CNN을 사용하여 특징벡터를 추출하였고, 다른 학습 없이 특징벡터의 거리를 사용하여 people re-id를 하였다. VGG 넷[1]은 Large Scale Visual Recognition Challenge 2014 (ILSVRC2014)의 Classification and localization 부분에서 좋은 성능을 보였다. 우리가 사용한 VGG-S 넷[2]은 VGG 넷의 버전 중 하나이다.

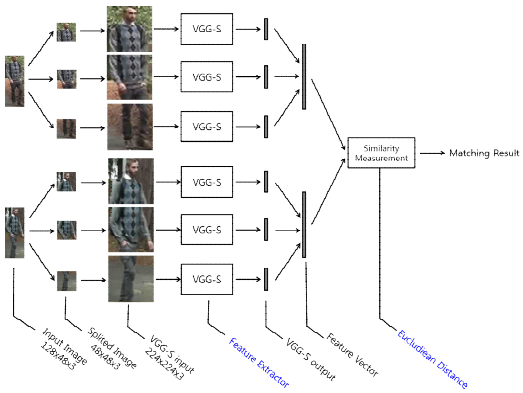
Ⅱ. 제안하는 시스템

People re-id에서 우리는 ImageNet 데이터 셋으로 사전학습된 VGG-s 넷을 사용하여 특징을 추출하였다. 이 과정에서 어떠한 학습도 하지 않았다. VGG-s 넷은 순순하게 특징 추출기로 동작하였다. VGG-s 넷은 객체분류를 목적으로 설계되었다. 반면 우리의 응용은 People re-id로 객체분류와 다른 분야이다. VGG-s에서 각 클래스의 확률을 만드는 마지막 Softmax 단을 제외하고 7번째 Fully Connected층(FC7)의 출력, 6번째 Fully Connected층(FC6)의 출력, 5번째 Pooling층(POOL5)의 출력을 각각의 특징벡터로 사용하여 비교하였다. 두 이미지의 유사성 측정으로는 추출된 특징벡터간의 Euclidean distance를 사용하였다.

VGG-s 넷의 입력 이미지 사이즈 224*224*3와 VIPeR 데이터 셋의 이미지 사이즈 128*48*3를 서로 맞춰줘야 한다. 우리는 다음 방법으로 이미지 사이즈를 맞췄다. 테스트 이미지를 중첩되게 세로로 3개의 48*48*3 이미지로 나눈 후 Spline Interpolation으로 확대시킨 것이다. 이것은 그림1에서 볼 수 있다.

Ⅲ. 실험 및 결과

테스트 프로그램은 Theano 라이브러리와 Lasagne 라이브러리를 사용하여 파이썬으로 작성되었다. pre-train된 VGG_S 모델은 이곳[3]에서 다운받았다.



▶▶ 그림 1. 제안하는 시스템 구성

VGG-s 넷을 사전학습한 데이터 셋은 ImageNet이다. 이것은 객체분류를 위한 데이터 셋으로 1,000개의 클래스와 456,567개의 이미지로 구성된다. People re-id를 위한 데이터 셋으로는 VIPeR 데이터 셋[4]을 사용하였다. VIPeR 데이터 셋은 632쌍의 이미지로 구성되었으며, 각 이미지의 사이즈는 128*48*3이다.

People re-id의 성능 측정을 위해서 Expected Rank, AUC를 사용하였다. Expected Rank는 매칭된다고 판단한 사람의 등수에 대한 기대값이며, 작을수록 좋다. AUC는 Rank k를 그래프로 그렸을 때 그래프 아랫 부분의 넓이이며, 0과 1사이의 값이다. 이것은 1에 가까울 수록 좋다. Rank k는 매칭된다고 판단한 사람이 등수 k안에 들어갈 확률을 뜻하며 클수록 좋다. 랜덤한 선택을 했을 때 Rank 1은 1/number of data, Expected Rank는 number of data/2, AUC는 0.5이다.

성능 비교를 위하여 SDALF[5], ELF[6]과 비교하였다. 이것들은 최신의 연구는 아니다. 하지만 우리의 실험은 학습을 하지 않았기 때문에 좋은 비교가 될 것이라고 생각한다. 다만 성능 평가 메서드는 테스트 집합의 개수에 영향을 받기 때문에 다른 비교 방법과 같이 316개의 데이터로 FC7에 대해서만 실험하였다. 표1은 실험 결과를 나타낸다. ER은 Expected Rank를 의미한다.

우리는 학습을 하지 않고, 다른 응용으로 사전학습한 네트워크를 사용했지만, 그럼에도 ELF와 비슷한 성능을 보여줬다. 이것은 CNN이 다른 응용을 목적으로 학습하여도 영상에서 충분히 좋은 특징을 추출하는 것을 생각된다. POOL5, FC6, FC7에서의 특징 추출 실험에서는 네트워크가 깊을수록 좋은 특징을 추출하는 모습을 확인하였다.

Ⅳ. 결론

최근 객체인식 분야에서는 CNN이 주목받고 있다. CNN의 특징 중 하나는 입력이미지로 부터 특징 추출 방법을 스스로 학습한다는 것이다. 전통적은 객체인식 방법에서는 hand-written feature extractor를 사용하지만, CNN은 스스로가 특징을 추출한다. 우리는 객체분류를 위한 ImageNet 데이터 셋으로 사전학습된 VGG-s 넷을 사용하여 특징을 추출하였다. 이 과정에서 어떠한 학습도 하지 않았다. 반면 우리의 응용은 people re-id로 서로 다른 분야이다.

표 1. 실험 결과

	ER	AUC
SDALF	44.7	0.86
ELF	83.2	0.74
Ours FC7 (with half data)	79.9	0.75
Ours POOL5	264.5	0.59
Ours FC6	163.1	0.74
Ours FC7	147.7	0.77

우리는 학습을 하지 않고, 다른 응용으로 사전학습 네트워크를 사용했지만, 그럼에도 ELF와 비슷한 성능을 보여줬다. 이것은 CNN이 다른 응용을 목적으로 학습하여도 영상에서 충분히 좋은 특징을 뽑아낸다는 것을 의미한다고 생각된다. 앞으로 우리는 다른 CNN에 모델에 대해서 이 실험을 반복하고, 학습 알고리즘을 적용해 볼 예정이다.

사사

이 논문은 한국연구재단 BK21 플러스 사업의 지원과 2013년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임. (No.2013R1A1A2013778)

■ 참 고 문 헌 ■

[1] Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", 2015, arXiv:1409.1556

[2] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, "Return of the Devil in the Details: Delving Deep into Convolutional Nets", British Machine Vision Conference, 2014

[3] Lasagne/Recipes/examples/ImageNet Pretrained Network(VGG_S), https://github.com/Lasagne/Recipes/blob/master/examples/ImageNet%20Pretrained%20Network%20(VGG_S).ipynb, (Access 2016.04.28.)

[4] D. Gray, and H. Tao, "Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features," in Proc. European Conference on Computer Vision (ECCV), 2008.

[5] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In IEEE Conference on Computer Vision and Pattern Recognition. 2010.

[6] Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European Conference on Computer Vision. Marseille, France (2008)