Web-Videos를 사용한 Supervised Learning Framework

*나성원 **이예지 ***윤경로
건국대학교
*securityin4@naver.com

Supervised learning framework using Web-Videos

*Na, Seong-Won **Lee, Ye-Gi ***Yoon, Kyoung-ro
Konkuk University

요약

본 논문에서는 비디오 데이터를 이용한 감독 학습 프레임 워크를 제안한다. 최근 Deep Convolutional Neural Networks의 성공으로 많은 분야에서 사용되고 있다. DCNNs 모델 성능의 중요한 요소 중 하나는 Large-cale Dataset을 구축하는 것으로 Small-scale Dataset으로 모델을 학습한다면 과적합 및 일반화 오류를 해결하기 어렵다. 이러한 문제점을 해결하는 방법으로 이미지 왜곡을 통한 데이터 셋을 증가 또는 Dropout 기법 등을 사용하였지만 원본 데이터가 적은 경우에는 모델이 일반화 능력을 갖기 어렵다. 따라서 본 논문에서는 이러한 문제점을 보완하고자 Web으로부터 얻은 비디오에서 해당 Class와 관련된 프레임들을 추출하여 보다 쉽게 데이터 셋을 확장하고, 모델의 성능을 향상 시키는 방법을 제안한다.

1. 서론

최근 Deep Convolutional Neural Networks(DCNNs)의 성공으로 다양한 분야에서 적용하고, 성능을 높이기 위해 연구되고 있다. DCNNs의 성능 향상 요인 중 중요한 한 가지는 Large-scale Dataset의 지원 여부이다. 하지만 대부분의 분야에서는 적당한 데이터 셋이 존재하지 않으며, 데이터 셋을 구성하는 것 또한 힘들다. 기본적으로 하나의 클래스를 확장할 경우 몇 백에서 몇 천개의 이미지를 Web이나 직접 촬영을 통해 얻어야 하는데 이는 시간과 노력, 비용적인 면이 상당히 크기 때문이다. 이전 연구들에서는 데이터를 증가시키는 방법으로 이미지 왜곡(Rotation, Flip, Shift, Rescale)을 하나의 전처리 과정으로 추가하지만, 원본 학습 데이터 셋이 부족한 상태로 모델을 훈련하게 되면 과적합 및 일반화 능력이 떨어지게 되기 때문에 효과가 없다.

따라서 본 논문에서는 이러한 문제점을 극복하고자 Web-Video를 사용한 새로운 프레임워크를 제안 한다. 이전에 연구된 방법 중 Web-Video 데이터를 직접적으로 훈련시킨 것[1]과 다르게 본 논문에서는 비디오 내에서 클래스와 관련 있는 프레임들만 선택한다. 하나의 비디오 내에는 클래스 이름으로 검색 했다 하더라도 관련 없는 내용을 많이 포함하고 있어 작은 데이터 셋일수록 모델 학습에 있어 치명적일 수 있기 때문에 이 작업은 중요하다.

관련 프레임을 선택하기 위해서 본 논문에서는 Deep Feature를 사용한다. Deep Feature는 DCNNs의 마지막 콘볼루션 레이어로부터 추출한 Feature로써 원본 트레이닝 셋 중에 임의의 한 이미지와 비디오의 프레임들에서 각각 추출하여 코사인 유사도를 측정해 임계 값 보다 높을 시 유사한 프레임으로 선택 된다. 이렇게 관련 있는 프레임들만 추출하고 난 뒤 비디오 데이터의 특징 중 하나인 유사 장면이 연속적으로 이어지는 것을 이용하여 추가적으로 프레임을 더 획득 하였다.

이렇게 획득 한 비디오 프레임들은 배경을 제거하는 단계를 추가적으로 실시하여 훈련 데이터로 사용하였다. 실험 시 Class Activation Mapping(CAM)[2]을 통해 확인한 결과 비디오 프레임 데이터가 많아질수록 똑같은 배경이 계속해서 나타나기 때문에 오브젝트에 집중하지 못하고 넓게 퍼져서 학습되는 것을 확인 하였다. 이러한 과정을 통해 얻은 데이터들을 실험하기 위해 사전 훈련된 모델을 사용하고, 3개의 레이어를 추가하여 실험 하였다. 논문의 구성은 본론에서 구체적인 프레임워크를 설명하고, 네트워크에 대해 짧게 설명하며, 그 후 실험결과를 분석하고 결론을 논의 한다.

2. 본론

2.1 프레임워크

그림 1은 본 논문에서 제안하는 전체 프레임 워크이다. 본 논문에서는 먼저 목표 클래스 이름으로 검색 된 비디오에서 관련 있는 프레임들을 추출하기 위해 해당 데이터 셋 중 하나의 이미지와 유사도 측정을 한다. 이를 위해 이미지를 벡터로 변환하게 되는데 여기서는 f_v로 표현 하였다. 벡터 변환은 사전 훈련 된 DCNNs 모델을 사용하며 마지막 콘볼루션 레이어로부터 특징 벡터를 출력 받으며, 유사도 측정 함수 f_s의 입력으로 사용 된다. f_s는 두 벡터 사이의 코사인 유사도를 측정하여 0에서 1사이의 값을 반환하게 되는데 특정 임계값을 설정하고, 임계값을 넘어가게 되면 관련된 프레임으로 선택 된다. f_s로부터 추출된 프레임은 비디오 데이터의 특성상 유사한 프레임이 연속적으로 연결되기 때문에 그 이후 4개의 프레임을 추가로 추출한다. 여기서 4개로 지정한 이유는 데이터 셋 자체가 작기 때문에 더 많은 프레임을 추출 했을 시 학습 결과가 좋지 않았기 때문이다. 이렇게 추출된 관련 프레임들은 같은 배경이 연속적으로 등장하기 때문에 본 논문에서는 학습시키고자 하는 오브젝트에 집중할 수 있도록 배경과 오브젝트

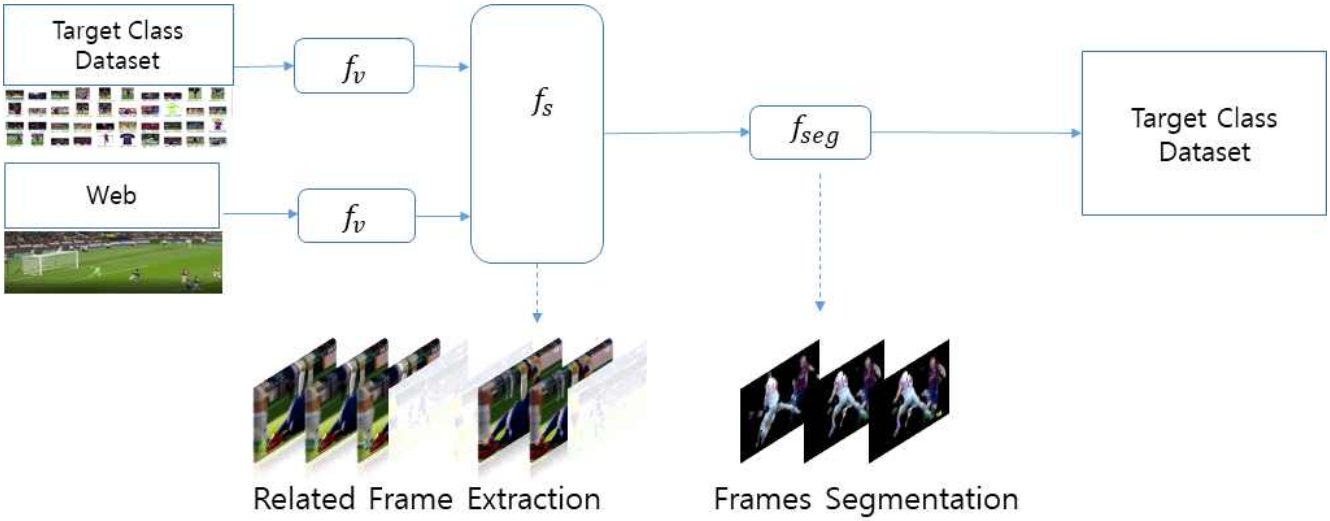

그림 1. 제안한 전체 프레임워크

를 분할하는 과정을 추가 하였다. f_{seg}는 위에서 말한 배경과 오브젝트를 분할하는 함수로 배경은 제거 된 후 이미지로 저장되어 해당 클래스의 훈련 데이터로 사용 된다. 하나의 검색 된 비디오는 이 과정을 여러 번 반복하게 되는데 이는 비교될 원본 훈련 데이터가 무작위로 선택 될 때 각 이미지가 포함하고 있는 내용이 다르기 때문에 유사도 측정 시 더 다양한 내용의 관련 프레임을 추출할 수 있으며 모델 일반화 성능에 도움을 줄 것으로 생각한다.

2.2 네트워크 구성

네트워크는 먼저 Imagenet[3]으로 사전 훈련된 Resnet모델을 기본 구조로 사용하였고, 마지막 Softmax layer를 제거하고, 추가적으로 Global Average Pooling Layer 1개와 Fully Connected Layer 2개를 붙였다. 3개의 클래스에 대해 실험을 진행하였고, 기본 구조에서 1000개의 클래스와 비교해 적은 수이기 때문에 차원수를 줄이고자 Fully Connected Layer를 512, 256으로 설정하여 네트워크를 구성 했다. GAP Layer는 이미지 분류 시 근거가 되었던 부분을 시각화 할 수 있는 Class Activation Mapping(CAM)을 확인하기 위해 추가 하였다.

3. 실험 및 결과

실험은 같은 네트워크를 사용하였고, 데이터 셋은 원본 데이터 셋, 제안한 프레임워크로 획득한 데이터 셋으로 구성하여 모델을 학습하고 결과를 도출 했다. 원본 데이터 셋의 경우 클래스 당 120장으로 구성 되어 있고, 추가로 획득한 데이터 셋은 클래스 당 360장으로 구성하였다.

그림 2는 각각의 데이터 셋으로 학습 시 측정된 정확도와 손실에 대한 값을 그래프로 나타낸 것이다. 두 그래프를 비교해보면 제안한 프레임워크로 측정한 지표가 원본 데이터 셋보다 Train, Validation Set에 대해 정확도가 5~10%정도 향상 되었으며, 학습 그래프 또한 더 부드럽게 형성되어 비교적 안정적으로 수렴한 것을 볼 수 있다.

그림 3은 이미지 분류 시 해당 라벨에 대해 상응하는 부분을 시각화 해주는 CAM을 보여주고 있다. 먼저 원본 데이터 셋으로 학습 후 CAM을 생성한 이미지를 보면 객체와 배경 모두에 대해 고르게 강조가 되어 있는 것을 볼 수 있다. 본 논문에서는 제안한 프레임워크로 얻은 데이터 셋으로 훈련 된 모델은 그림 3에서 볼 수 있듯이 객체에 집중 되어 강조 된 것을 볼 수 있다.

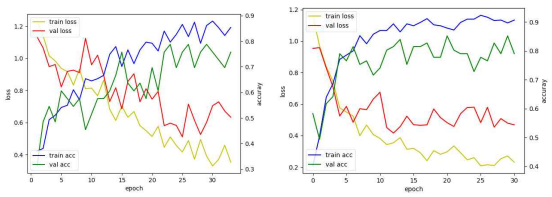

그림 2. 각 데이터 셋에 대한 정확도, 손실 측정 그래프 (왼쪽 원본, 오른쪽 제안한 프레임워크)

그림 3. 각 모델의 테스트 이미지에 대한 CAM 시각화 (왼쪽 원본, 오른쪽 제안한 프레임워크)

4. 결론

본 논문에서는 비디오 데이터를 사용한 감독 학습 프레임 워크를 제안 한다. 제안된 프레임 워크는 다양한 분야에서 DCNNs의 도입에 중요한 요소 중 하나인 Data를 증가시키는 방법으로 웹에서 해당 클래스로 검색된 비디오 데이터를 사용하여 관련 프레임을 획득하고, 획득한 프레임들에 대해 잡음을 제거하여 오브젝트에 집중할 수 있도록 가공한 후 데이터 셋으로 사용 하였다. 실험 결과에서 볼 수 있듯이 모델이 안정적으로 훈련되었고, 정확도도 향상 되었다.

감사의 글

이 논문은 2019년 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(S201901S00055, UHD 방송콘텐츠 기반 지능형 Dynamic Media 생성, 분배 및 소비 기술 개발)

참고문헌

[1] Tokmakov, Pavel, Karteek Alahari, and Cordelia Schmid. "Learning semantic segmentation with weakly-annotated videos." ECCV. Vol. 1. 2016.

[2] Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[3] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.