Bigdata Analytics

최종 보고서

과목명	빅데이터분석론	
담당교수님	황지영 교수님	
학부	전자전기컴퓨터	
조 이름	3 조	
학번	2018711164	2017711778
이름	최은빈	박지윤
제출날짜	2018. 06. 13	

1. 서론

I. 연구선정 배경 및 필요성

[프로젝트의 최종목표] 사용자가 필요로 하는 드론 영상을 쉽게 찾을 수 있도록 드론 영상을 분석하여 해당 영상과 관련된 키워드를 추출하는 기술 개발

드론은 이동의 제약이 없고 사용자가 원거리에서 실시간으로 영상을 받아볼 수 있다는 특징이 있기 때문에 사람이 직접 카메라를 이용하는 기존 영상 촬영방식에서 벗어나, 보다 다양한 구도의 영상을 획득하는 것이 가능하다. 이를 통해 기존에 촬영이 어려웠던 장소나 쉽게 이동할 수 없는 작업환경에서도 많이 활용된다. 초기에는 군사용으로 형성되었던 드론 시장이 의학, 기상, 과학, 예술분야 등 다양한 전문 분야 및 민간 상업용과 레저용으로 확대되면서 일상생활에서의 다양한 드론 영상이 선보여지고 있으며, 이에 따른 데이터 양이 급격히 증가되는 추세이다. 증가되는 드론 영상 데이터 양에 비해 실질적으로 유의미하게 사용되는 드론 영상 데이터는 매우 한정적이다. 이러한 한계점을 극복할 수 있는 드론 영상 분석 기술이 필요하다.

II. 연구 목표

본 프로젝트에서는 사용자가 필요로 하는 드론 영상 데이터를 쉽게 찾아내기 위하여 드론 영상에 담긴 정보를 분석하고 해당 영상과 관련된 키워드를 추출하는 기술을 개발하고자 한다.

III. 사전 연구

비디오 캡션의 이전 연구는 메타데이터를 가지고 비디오를 tagging[1]하는 것과 검색 작업을 위해 캡션과 비디오를 클러스터링[8,11,16] 하는 것이다. 문장 설명[6,9,13]을 생성하는 것에 있어 이전 방법에서는 2 개의 단계 파이프라인을 사용했다. 첫번째는 의미론적 내용(주어, 동사, 목적어)을 확인하고 그 다음으로 템플릿에 기반한 문장을 생성하는 것이다. 이는 후보가 되는 물체, 행동, 장면을 식별하기 위해 개별의 classifiers 를 학습하는 것을 전형적으로 포함한다. 그들은 가장 가능성이 높은 내용(주어, 동사, 목적어, 장면)을 추정하기 위해 언어 모델과 함께 시각적 신뢰를 포함하는 확률론적 문법 모델을 사용한다. 이는 내용 생성과 surface realization 을 분리시키면서 문제를 간단화하기 위해서는 관련 객체 세트와 인식할 동작을 선택해야한다. 게다가, 문장을 생성할 때 템플릿 기반의 접근법은 인간 묘사에 사용되는 언어의 풍부함을 모델링하기에는 불충분하다. 예를 들어, 어떤 속성을 사용하고 어떻게 효과적으로 결합하여 좋은 설명을 생성할 것인가에 대한 문제이다. 본 연구에서 제안한 모델은 비디오를 제공된 문장으로 맵핑된 것을 학습시키고, 언어 모델을 시각적 특징을 조건으로 하여 학습하면서 문장 생성이 내용 식별과 분리되는 것을 피한다.

본 연구는 [4,15]에 있는 이미지 캡션 생성 모델로부터 영감을 얻었다. 그들의 모델은 첫번째로 CNN 을 통해 특징을 추출함으로써 이미지의 고정된 길이 벡터 표현(optical view)을 생성하는 것이다. 다음 스텝은 이 벡터를 이미지의 설명을 구성하는 단어 시퀀스로 디코딩하는 방법을 학습한다. RNN 은 시퀀스를 디코딩하는 원리에 사용되었지만 결과적으로 long-term dependencies 는 열등한 성능을 이끌었고, 이를 완화시키기 위해 LSTM 모델은 시퀀스 디코더로 사용되었고, 그들은 long-range dependencies 를 학습하기에 적합했다. 게다가 본 연구에서는 가변적인 길이의 비디오를 인풋으로 사용하기 때문에 [12]의 언어 번역 모델에 따라 시퀀스 변환기에 대한 시퀀스로서 LSTM 을 사용한다. [14]에서 LSTMs 을 사용하여 개별 프레임의 표현을 풀링하여 비디오 설명을 생성한다. 그들의 기술은 비디오에서 프레임에 대해 CNN 특징을 추출하고, 그때 그 결과를 mean-pooling 하여 전체 비디오를 나타내는 단일 피쳐 벡터를 얻는다. 그때 LSTM 을 이 벡터에 기반을 둔 설명을 생성하기 위해 시퀀스 디코더로서 사용한다. 이 접근법의 주된 결점은 이러한 설명은 완전히 비디오 프레임의 순서를 무시하는 일이고 이는 시간적인 정보를 이용하는 것을 실패한다. [4]에 접근법은 LSTM 을 이용하여 비디오 설명을 생성한다; 그러나, 그들은 CFS 를 사용하여 활동, 객체, 도구 및 위치 지정의 의미 튜플을 얻은 다음, LSTM 을 사용하여 이 튜플을 문장으로 변환하는 2 단계 접근 방법을 사용한다. 게다가, [4]의 모델은 비디오를 변환하는 제한된 영역에서 적용되는 반면, 본 연구에서는 "in the wild" 비디오에 대한 설명을 생성하는 것을 목표로 한다.

본 [17]에서 접근법은 또한 [14]의 제한을 두 가지 방법으로 나눈다. 첫째, 그들은 시공간 운동 특징을 포함하는 3-D convnet 모델을 사용한다. 이 피쳐를 얻기 위해서는 그들은 동영상이 고정된 볼륨(너비, 높이, 시간)이라고 가정한다. 그들은 겹치지 않는 직육면체 위에 조밀한 궤도 특징(HoG, HoF, MBH)을 추출하고, 이들을 연결하여 입력을 형성한다.

3-D convnet 은 동작 인식을 위해 비디오 데이트셋에 대한 사전 학습을 한다. 둘째, [14]에서와 같이 모든 프레임의 피쳐를 균일하게 가중치시키기보다는 이전 워드 입력에 대해 불균형하게 컨디셔닝 된 프레임 피처에 무게를 가하는 학습 매커니즘을 포함한다. 3-D convnet 은 제한된 성능 향상을 제공하지만, 주의 모델과 함께 성능을 현저히 향상시킨다. 본 연구에서는 LSTM 을 사용하여 비디오 프레임의 시퀀스를 문장 설명으로 생성하기에 충분한 분산 벡터 표현으로 인코딩함으로써, 시간 정보를 사용하는 보다 간단한 방법을 제안한다. 따라서 시퀀스 모델에 대한 직접 시퀀스에는 명시적인 주의 메커니즘이 필요하지 않다.
또다른 최근 프로젝트는 이전 프레임을 인코딩함으로써 특징 프레임 시퀀스를 예측하기 위해 LSTMs 를 사용한다. 그들의 모델은 [12]에서 언어 번역 모델과 더 비슷하다. 그것은 하나의 LSTM 을 고정된 표현으로 인풋 텍스트를 인코딩하기 위해 사용한다. 그리고 이를 디코딩하기 위해 또다른 LSTM 을 다른 언어에 사용한다. 대조적으로 본 연구에서는 인코딩과 디코딩 둘다 이것이 제공된 인풋을 기반하여 학습할 수 있는 단일의 LSTM 을 사용한다. 이는 인코딩과 디코딩 사이에 가중치를 공유하기 위해 LSTM 을 사용한다.

2. 본론

I. 연구 내용

우리는 비디오 설명에 대해 sequence to sequence 모델을 제안한다. 이는 입력은 비디오 프레임의 시퀀스 $(x_1,...,x_n)$ 이고, 출력은 단어의 시퀀스 $(y_1,...,y_m)$ 이다. 자연스럽게 입력과 출력은 모두 가변적이고 잠재적으로 길이가 다르다. 본 연구에서는 전형적으로 단어의 수보다 프레임의 수가 많다.

우리의 모델은 입력 시퀀스 $(x_1,...,x_n)$ 가 주어질 때, 출력 시퀀스 $(y_1,...,y_m)$ 의 조건부 확률을 추측한다. 예를 들어,

$$p(y_1,\ldots,y_m \mid x_1,\ldots,x_n) \quad (1)$$

이 문제는 자연어들 사이에서 기계 번역과 유사한 것이다. 이는 입력 언어에서 단어의 시퀀스가 출력 언어에서 단어의 시퀀스로 번역된다. 최근에 [3,12]는 어떻게 이러한 sequence to sequence 문제를 LSTM RNN 을 사용하여 효율적으로 해결할 수 있을지에 대해 보여주었다. 우리는 이러한 패러다임을 비디오 설명에서 RNN 기반의 방법을 상당히 간략화하면서 비디오 프레임의 시퀀스로 구성한 입력으로 연장한다. 다음에서, 우리는 비디오와 문장에 대해 입력과 출력 표현뿐만 아니라 우리의 모델과 구조를 설명한다.

- 시퀀스 모델링을 위한 LSTMs

가변 길이의 입력과 출력을 다루는 주된 아이디어는 첫째로 프레임의 입력 시퀀스를 한번에 latent vector 표현을 사용하여 비디오를 인코딩하는 것이다. 그리고 한번의 한 단어씩 표현에서 문장으로 디코딩한다.

첫째로, 원래 [7]에서 제안된 LSTM 을 생각해보면 [18]에서 제안했던 LSTM unit 을 기반으로 Time step t 의 Input x_t 에 대해 LSTM 은 hidden/control state h_t 와 time t 까지 cell 이 관찰해온 모든 것을 인코딩하는 memory cell state c_t 를 계산한다.

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\ g_t &= \phi(W_{xg}x_t + W_{hg}h_{t-1} + b_g) \\ c_t &= f_t \odot c_{t-1} + i_i \odot g_t \\ h_t &= o_t \odot \phi(c_t) \end{aligned} \quad (2)$$

여기서 sigmoidal non-linearity σ , hyperbolic tangent non-linearity ϕ , element-wise product with the gate value $\odot$, weight matrice W_{ij} , biases b_j 은 학습된 파라미터들이다.

그래서 인코딩 단계에서는 입력 시퀀스 $X(x_1,...,x_n)$ 가 주어지면, LSTM 은 히든 시퀀스 $h_1,...,h_n$ 를 계산한다. 디코딩하는 동안 입력 시퀀스 X 가 주어질 때, 출력 시퀀스 $Y(y_1,...,y_m)$ 에 대한 분포를 $p(Y \mid X)$ 로 정의한다.

$$p(y_1,\ldots,y_m \mid x_1,\ldots,x_n) = \prod_{t=1}^{m} p(y_t \mid h_{n+t-1}, y_{t-1}) \qquad (3)$$

여기서 $p(y_t \mid h_{n+t})$ 의 분포는 단어 안에 있는 모든 단어에 대해 $softmax$ 로 주어진다(Equation 5). Equation 2 에서 순환적인 성질에 기반하여 h_{n+t-1}, y_{t-1} 으로 h_{n+t} 이 얻어진다는 것에 주목할 수 있다.

- Squence to Squence Video to Text

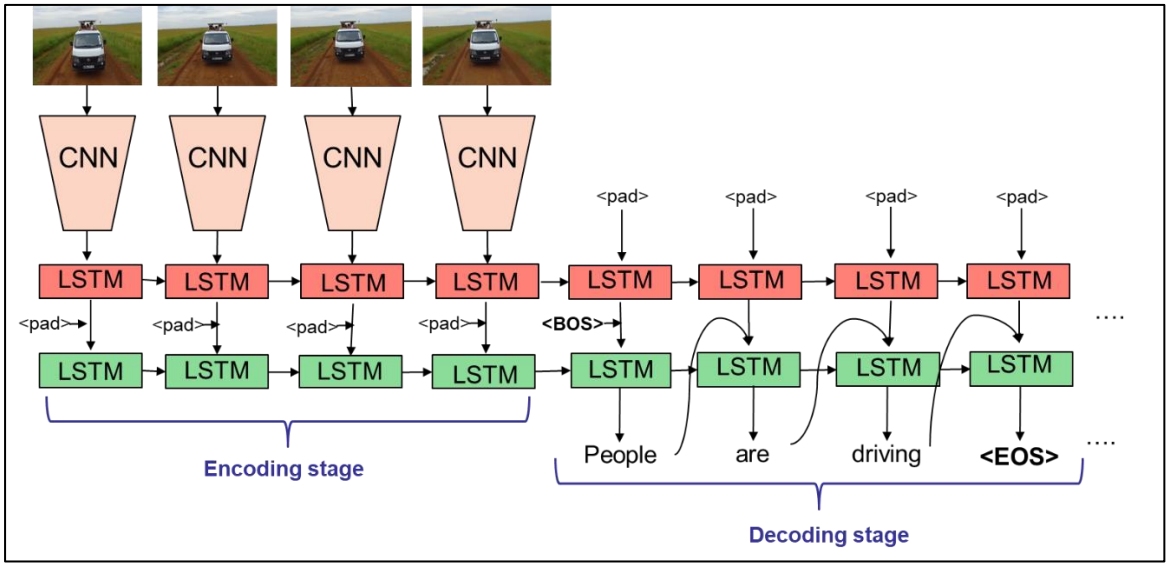


Figure 1. Squence to Squence Video to Text 구조

[3,12]는 첫째로 인풋 시퀀스를 고정된 길이 벡터로 인코딩하고, 그때 벡터를 아웃풋의 시퀀스로 매핑하기 위해 또다른 LSTM 을 사용하는 반면에, 우리는 인코딩과 디코딩 단계 모두 단일의 LSTM 을 사용한다. 이는 인코딩과 디코딩 단계 사이에 파라미터를 공유하는 것을 가능하게 한다.

우리의 모델은 각각 1000 개의 히든 유닛과 함께 2 개의 LSTM 의 스택 형태이다. Figure 1 는 시간이 가면서 펼쳐지는 LSTM 스택을 보여준다. 두개의 LSTMs 는 함께 쌓여져 있을 때, 우리의 경우에는 첫번째 LSTM(빨간색) 층에서 숨겨진 표현(h_t)은 두번째 LSTM(초록색)의 인풋(x_t)으로 제공된다. 위층의 LSTM 은 시각적 프레임 시퀀스를 모델링하는데 사용되고, 다음 층은 아웃풋 워드 시퀀스를 모델링하는데 사용된다.

트레이닝과 추론 처음에 몇차례 스텝에서, 두번째 LSTM 층은 히든 레이어의 결과값을 받고 NULL 로 패딩된 입력 단어와 이를 연결하는 동안, 위층의 LSTM 은 (Figure 1 에서 빨간색으로 칠해진) 프레임들의 시퀀스를 받고 그들을 인코딩한다. LSTMs 이 인코딩 되는 단계 동안 손실이 없다. 비디오 클립에서 모든 프레임이 쓰여지면, 두번째 LSTM 층은 문장의 시작 <BOS> 태그를 받는다. 그것은 숨겨진 표현을 단어의 시퀀스로 디코딩하는 것을 시작한다. 디코딩 단계에서 트레이닝하는 동안에 모델은 비디오 프레임 시퀀스의 숨겨진 표현과 이전 단어가 주어질 때, 예측된 결과 문장의 log-likelihood 를 최대화한다. Equation 3 에서 모델의 파라미터 θ 와 아웃풋 문장 $Y(y_1,...,y_m)$ 에 대해, 다음과 같이 공식화한다:

$$\theta^* = \arg\max_{\theta} \sum_{t=1}^{m} \log p(y_t \mid h_{n+t-1}, y_{t-1}; \theta) \qquad (4)$$

이 log-likehood 는 stochastic gradient decent 를 사용하여 트레이닝 데이터셋 전부에 대해 최적화되어있다. 손실은 오직 LSTM 이 디코딩하기 위해 학습될 때 계산된다. 이러한 손실이 시간에 따라 전파됨에 따라 LSTM 은 입력

시퀀스의 숨겨진 상태 표현(h_n)을 생성하는 것을 학습한다. 두번째 LSTM 층의 출력(z_t)은 방출된 단어(y)를 얻는데 사용된다. 우리는 Vocabulary V 에 있는 단어 y'에 대해 확률 분포를 얻어내기 위해 $softmax$ 을 사용한다.

$$p(y \mid z_t) = \frac{\exp(W_y z_t)}{\sum_{y' \in V} \exp(W_{y'} z_t)} \qquad (5)$$

디코딩하는 동안 우리는 첫번째 LSTM 층에서 시각적 프레임 표현은 간단히 패딩 입력 역할을 하는 zero 벡터라는 것을 주목한다. 우리는 문장의 길이가 다양한 시퀀스에 대해 분포를 distribution 을 정의하는 모델을 가능하게 하기 때문에 각각 문장을 종결하기 위해 명백한 end-of-sentence 태그(<EOS>)를 요구한다. 테스트에서는 디코딩 스텝 동안 $softmax$ (Equaiton 5) 후에 이것이 <EOS> 토큰을 방출할 때까지 최대 확률을 가지는 단어 y_t 를 선택한다.

- 비디오와 텍스트 표현

RGB 프레임 이전 LSTM 기반의 이미지 캡셔닝[4,15]과 video-to-text 접근법[14,17]과 비슷하게, 우리는 입력 이미지에 CNN 을 적용하고, LSTM 유닛에 입력으로 첫번째 층의 출력을 제공한다. 이러한 연구에서, 우리는 Caffe Reference Net 과 16 층 VGG 모델을 fc7 층의 출력을 이용했다. ImageNet 데이터 셋의 1.2M 이미지 ILSVRC-2012 객체 분류 하위 집합에 사전 학습되었던 CNNs 을 사용하며 Caffe ModelZoo 를 통해 공개적으로 사용할 수 있다. 각각의 입력 비디오 프레임은 256x256 으로 스케일되고, 임의의 227x227 영역으로 자른다. 그리고 CNN 을 이용하여 특징을 추출한다. 기존 구조에서 가장 마지막의 fully-connected 층을 제거하고 500 차원 공간에 특징에 대한 새로운 선형 임베딩을 학습하고, 더 낮은 차원의 특징은 입력 (x_t)을 첫번째 LSTM 층에서 형성한다. 임베딩의 가중치는 학습하는 동안 LSTM 층과 함께 공동으로 학습된다.

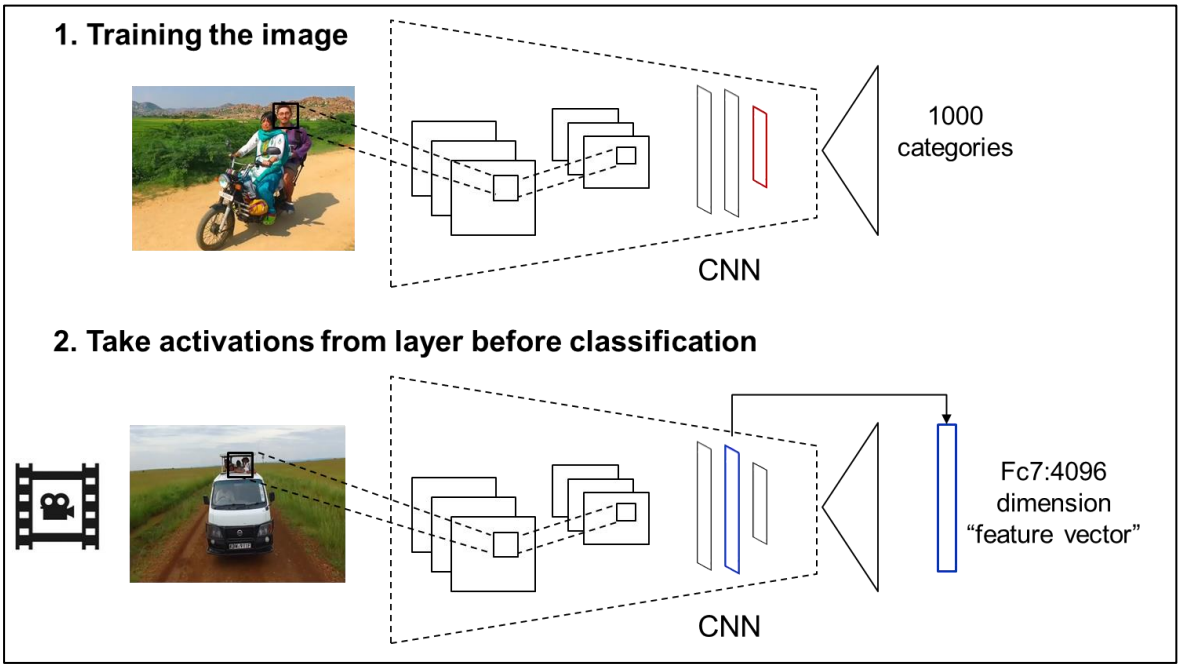


Figure 2. RGB 프레임 구조

Optical flow 원본 이미지 (RGB) 프레임으로부터 CNN 결과뿐만 아니라, 우리의 구조에 입력 시퀀스로써 옵티컬 플로우 측정을 통합한다. 다른 연구[10,4]에서는 LSTMs 에 optical flow 정보를 통합하는 것은 활동 분류를 증진한다. 우리의 설명의 대부분은 활동을 주심으로 하기 때문에, 우리는 또한 비디오 설명에 이러한 옵션을 탐색한다. 우리는 [4,5]에서 이러한 접근법을 따르고 첫째로 optical flow 특징[2]을 추출한다. 우리는 [5]의 방법과 비슷하게 x 와 y 값을 중심으로 하고 flow 값이 0 과 255 사이에 두기 위해 스칼라 값을 곱하는 것으로 flow 이미지를 생성한다. 또한 flow 크기를 계산하여 이를 flow 이미지에 세번째 채널로서 추가한다. 우리가 제안한 결합 모델에서, 우리는 플로우와 RGB 특징을 통합하기 위한 shallow fusion 기술을 사용한다.

텍스트 입력 타겟 단어의 출력 시퀀스는 one-hot 벡터 인코딩을 사용하여 표현된다. 프레임 특징의 처리와 마찬가지로, 입력 데이터에 선형 변환을 적용하고 역전파를 통해 매개 변수를 학습하여 더 낮은 500 차원 공간에

단어를 포함한다. 첫번째 LSTM 레이어의 아웃풋와 함께 연결된 포함된 단어 벡터는 두번째 LSTM 레이어(Figure 1 에서 초록색으로 표시된)의 인풋으로 형성한다.

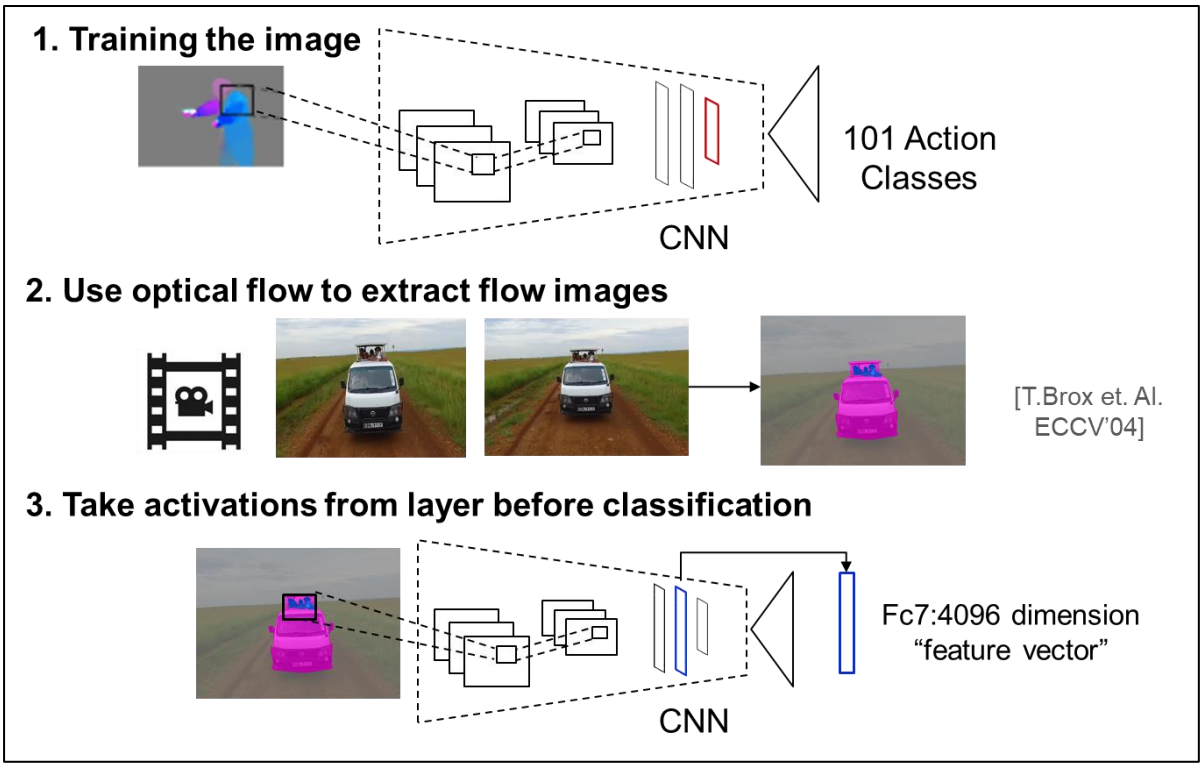


Figure 3. Optical flow 구조

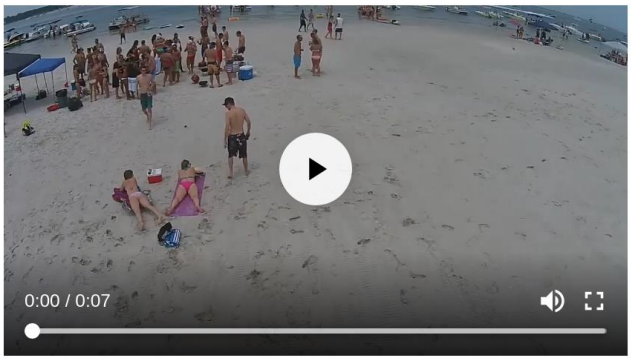
3. 결론

I. 연구 결과

본 프로젝트는 [19] 모델을 기반으로, YouTube 에서 수집한 5 초 이내의 1920 개의 드론 영상에 대해 1200 개의 영상에 대해서 training 을 수행, 100 개에 대해서는 validation, 그리고 나머지 620 개에 대해서는 test 을 수행하였다. 다음 표는 해당 프로젝트의 결과를 문장 예측이 정확하게 잘된 경우, 관련성은 있지만 정확하지는 않은 경우, 관련성도 없고 정확하지도 않은 경우로 분류하였다. 각각의 표 안의 이미지에는 비디오를 한 문장으로 나타내는 캡션과 이를 대표하는 단어를 추출하였다.

1) 문장 예측이 정확하게 잘된 경우

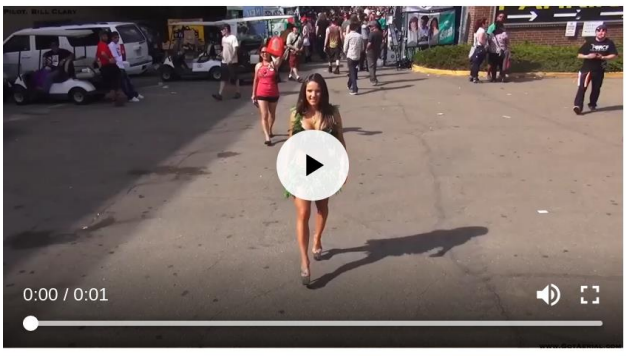
(1)



ID	#Words	Caption
beam (size 1)	9	A group of people are playing in the ocean.

Representative Word : ['playing', 'people', 'ocean', 'group']

(2)



ID	#Words	Caption
beam (size 1)	9	A group of people are walking on a road.

Representative Word : ['walking', 'road', 'people', 'group']

(3)

ID	#Words	Caption
beam (size 1)	8	A group of persons are doing a lake.

Representative Word : ['persons', 'lake', 'group']

(4)

ID	#Words	Caption
beam (size 1)	9	A group of men are walking on a road.

Representative Word : ['walking', 'road', 'men', 'group']

(5)

ID	#Words	Caption
beam (size 1)	9	A group of people are dancing in a stage.

Representative Word : ['stage', 'people', 'group', 'dancing']

(6)

ID	#Words	Caption
beam (size 1)	11	A group of people are riding a boat in the ocean.

Representative Word : ['riding', 'people', 'ocean', 'group', 'boat']

(7)

ID	#Words	Caption
beam (size 1)	6	A man is driving a car.

Representative Word : ['man', 'driving', 'car']

(8)

ID	#Words	Caption
beam (size 1)	6	A man is riding a bike.

Representative Word : ['riding', 'man', 'bike']

2) 관련성은 있지만 문장 예측이 정확하지 않은 경우

(1)

ID	#Words	Caption
beam (size 1)	6	A man is shooting a gun.

Representative Word : ['shooting', 'man', 'gun']

(2)

ID	#Words	Caption
beam (size 1)	7	Some people are dancing on a stage.

Representative Word : ['stage', 'people', 'dancing']

(3)

ID	#Words	Caption
beam (size 1)	6	A man is doing a snow.

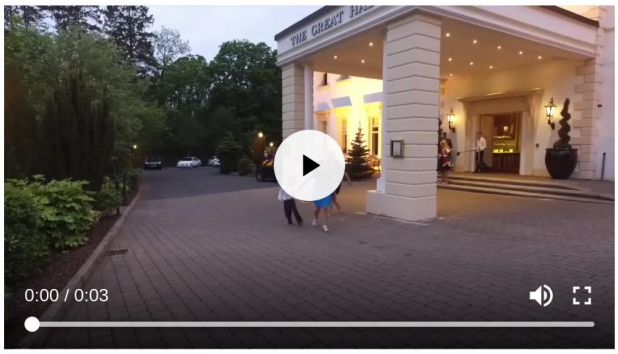
Representative Word : ['snow', 'man']

(4)

ID	#Words	Caption
beam (size 1)	6	A man is riding a hill.

Representative Word : ['riding', 'man', 'hill']

(5)



ID	#Words	Caption
beam (size 1)	8	A young man is walking in a building.

Representative Word : ['young man', 'walking', 'building']

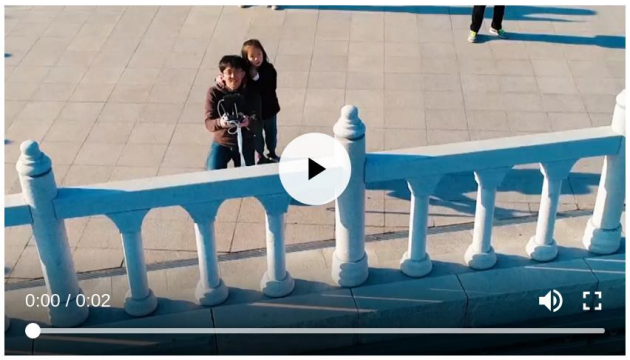
(6)

ID	#Words	Caption
beam (size 1)	6	A group of people are dancing.

Representative Word : ['people', 'group', 'dancing']

3) 관련성도 없고 문장 예측도 정확하지 않은 경우

(1)



ID	#Words	Caption
beam (size 1)	7	A woman is walking on a skateboard.

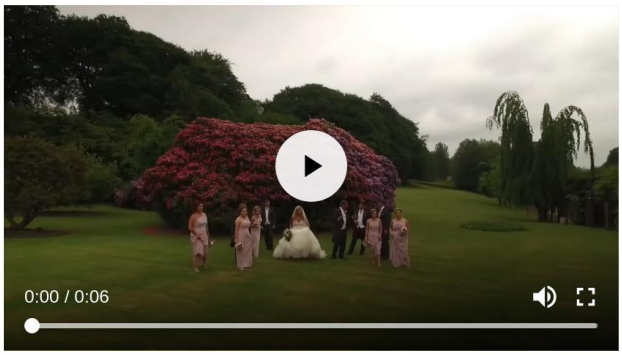
Representative Word : ['woman', 'walking', 'skateboard']

(2)

ID	#Words	Caption
beam (size 1)	7	A man is walking on a man.

Representative Word : ['walking', 'man']

(3)



ID	#Words	Caption
beam (size 1)	6	A man is riding a horse.

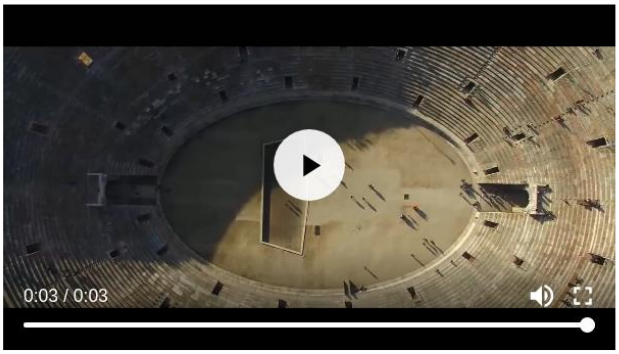
Representative Word : ['riding', 'man', 'horse']

(4)

ID	#Words	Caption
beam (size 1)	9	A group of a soccer game around a soccer.

Representative Word : ['soccer game around', 'soccer', 'group']

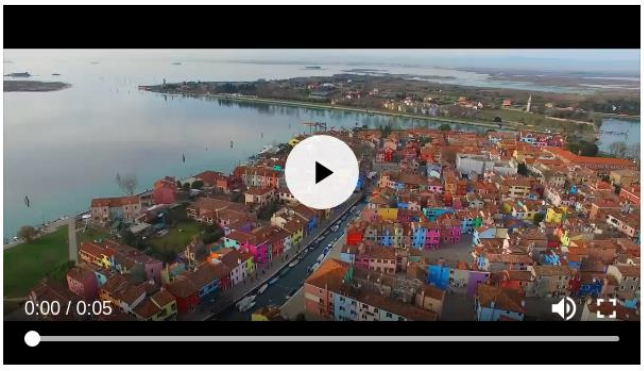
(5)



ID	#Words	Caption
beam (size 1)	6	A man is doing a music.

Representative Word : ['music', 'man']

(6)



ID	#Words	Caption
beam (size 1)	7	The people are riding in the road.

Representative Word : ['road', 'riding', 'people']

II. 문제점 및 해결방안

- 본 연구의 문제점은 객체가 뚜렷하게 드러나지 않는 비디오에 대해서는 캡션으로 표현하는 것이 어렵다는 것이다. 본 프로젝트에서 제안한 모델은 CNN 을 사용하여 시퀀스의 특징을 추출하고, 입력 시퀀스로서 optical flow 측정을 통합하기 때문에 활동 분류를 증진시킨다. 하지만 위의 분류 기준에서 3)의 (5), (6) 비디오에서는 객체가 뚜렷하게 드러나지 않고, 장면을 묘사하는 경우도 많기 때문에 비디오를 설명하기 위한 캡션 생성이 잘 이루어지지 않은 것을 볼 수 있다. 또한, 드론 영상에서 갑작스러운 화면 전환이 일어나거나 지속적으로 화면 흔들림이 발생하는 경우에는 이를 객체의 움직임으로 감지하는 경우도 있었다. 우리는 이러한 문제점을 해소하기 위해 다음 연구로는 드론의 흔들림 정보를 학습 모델에 함께 넣어주는 연구를 할 예정이다.

4. 참고자료

- [1] H. Aradhye, G. Toderici, and J. Yagnik. Video2text: Learning to annotate video content. In *ICDMW*, 2009.

- [2] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, pages 25–36, 2004.

- [3] K. Cho, B. van Merri¨enboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoderdecoder approaches. *arXiv:1409.1259*, 2014

- [4] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach,S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

- [5] G. Gkioxari and J. Malik. Finding action tubes. 2014.

- [6] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shoot recognition. In *ICCV*, 2013.

- [7] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8), 1997.

- [8] H. Huang, Y. Lu, F. Zhang, and S. Sun. A multi-modal clustering method for web videos. In *ISCTCS*. 2013.

- [9] N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko, and S. Guadarrama. Generating natural-language video descriptions using text-mined knowledge. In *AAAI*, July 2013.

- [10] J. Y. Ng, M. J. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. *CVPR*, 2015.

- [11] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, A. F. Smeaton, and G. Qu´eenot. TRECVID 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2012*, 2012.

- [12] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

- [13] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. J. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *COLING*, 2014.

- [14] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural

language using deep recurrent neural networks. In *NAACL*, 2015.

- [15] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *CVPR*, 2015.

- [16] S. Wei, Y. Zhao, Z. Zhu, and N. Liu. Multimodal fusion for video search reranking. *TKDE*, 2010.

- [17] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. *arXiv:1502.08029v4*, 2015.

- [18] W. Zaremba and I. Sutskever. Learning to execute. *arXiv:1410.4615*, 2014.

- [19] S. Venugopalan, M. Rohrbach, J. Donahue, T. Darrell, R. Mooney, K. Saenko. Sequence to Squence Video to Text

The IEEE International Conference on Computer Vision (ICCV) 2015.