creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

[UCI]I804:41014-200000376878

석사학위논문

스토리 영상 콘텐츠의 자동 이해를 위한 핵심 영역 중심의 이미지 캡셔닝 프레임워크

- Dense Captioning Framework Centered on Key Regions for Automatic Understanding of Video Content with Story -

지도교수 손 방 용

대진대학교 대학원

컴퓨터공학과 컴퓨터공학전공

소 진 수

2020년 12월

스토리 영상 콘텐츠의 자동 이해를 위한 핵심 영역 중심의 이미지 캡셔닝 프레임워크

- Dense Captioning Framework Centered on Key Regions for Automatic Understanding of Video Content with Story -

지도교수 손 방 용

이 논문을 공학석사 학위 논문으로 제출함

대진대학교 대학원

컴퓨터공학과 컴퓨터공학전공

소 진 수

2020년 12월

Collection @ daejin

소진수의 컴퓨터공학 석사학위 논문을 인준함

위 원 장 신 판 섭 (인)

위 원 정 종 진 (인)

위 원 손 방 용 (인)

대진대학교 대학원

2020년 12월

Collection @ daejin

국문요지

소 진 수

컴퓨터공학과 컴퓨터공학전공

대진대학교 대학원

딥러닝 분야에서는 이미지 및 영상을 시각적으로 분석하기 위해 이미지 분류, 객체 검출, 이미지 캡셔닝 등에 관한 연구들이 진행되었다. 이미지 캡셔닝 모델 중 하나인 밀집 캡셔닝 모델은 이미지 내 다양한 영역들을 추출하고 영역 수준의 캡션들을 생성하지만 영역 캡션에 대한 중요도를 제공하지 못하므로 이미지를 해석하는데 상대적으로 중요한 영역 캡션을 식별하기란 쉽지 않다. 또한, 스토리 콘텐츠인 영화 및 드라마 영상을 대상으로 한 연구가 매우 부족한 실정이다.

본 연구에서는 스토리 영상 콘텐츠 중 하나인 영화를 대상으로 영화 자동 이해를 위한 핵심 영역 중심의 이미지 캡셔닝 프레임워크를 제안한다. 제안 프레임워크는 DenseCap 모델을 기반으로 하며 캐릭터 식별 모듈, 핵심 영역 캡션 검출 알고리즘, 후처리기로 구성된다. 영화의 캐릭터 정보를 핵심 영역 캡션 검출과 캡션 개선에 활용하기 위해 HoG 알고리즘과 DBSCAN을 사용하여 캐릭터 얼굴 학습 데이터 세트를 생성하고, EfficientNet 모델을 학습하여 영화 장면 속 캐릭터를 식별하는 캐릭터 식별 모듈을 설계 및 구현하였다. 또한, 영역 캡션 중요도에 영향을 미치는 변수로 ‘영역 박스 신뢰도 점수’, ‘영역 면적’, ‘영역과 장면 중심 간 거리’, ‘객체 종류’ 4가지를 고려하는 핵심 영역 캡션 검출 알고리즘을 제안

Collection @ daejin

한다. 이후, 후처리기에서 동일 객체에 대한 영역 캡션을 하나로 통합하고, 영역 캡션을 영화에 적합한 형태로 개선하는 과정을 수행한다. 제안 프레임워크는 영화 장면의 중요한 영역 캡션을 효과적으로 식별하며, 기존 밀집 캡셔닝 모델보다 적은 수의 영역 캡션으로도 많은 정보를 제공하는 것을 확인하였다.

주요어 : 이미지 캡셔닝, 밀집 캡셔닝, 얼굴 인식, 영상 콘텐츠, 영화 이해

Collection @ daejin

목 차

Collection @ daejin

Collection @ daejin

그 림 목 차

Collection @ daejin

Collection @ daejin

표 목 차

Collection @ daejin

I. 서론

1. 연구 배경 및 목적

스마트폰, 고화질 카메라, CCTV, 드론 등으로부터 생성되는 이미지 및 영상 데이터가 기하급수적으로 증가하고 있으며, 넷플릭스, 유튜브, 인터넷 TV 등의 영상 콘텐츠가 전체 IP 트래픽에서 차지하는 비율이 2021년에는 82%를 차지할 것으로 예상된다[2]. 이러한 비정형 데이터를 시각적으로 분석하여 의미 있는 정보를 도출하고 활용할 수 있는 기술의 필요성이 대두되었다. 딥러닝 분야에서는 이미지 및 영상을 분석하기 위해 이미지를 몇 가지 레이블(label)로 예측하는 이미지 분류(image classification)부터 이미지 내 객체를 예측하고 레이블을 부여하는 객체 검출(object detection)과 이미지에 대한 설명을 자연어 형태로 생성하는 이미지 캡셔닝(image captioning)에 관한 연구들이 활발히 진행되었다. 최근에는 더욱 풍부한 캡션(caption)을 생성하기 위해 객체 검출 모델과 이미지 캡셔닝 모델을 통합하여 이미지 내 다양한 영역(region)들을 추출하고 영역 수준의 캡션을 생성하는 밀집 캡셔닝(dense captioning) 모델들이 제안되었다. 이러한 밀집 캡셔닝 모델들은 다양한 영역 캡션(region caption)을 생성하지만, 이미지를 해석하는데 상대적으로 중요한 영역 캡션을 식별하기 쉽지 않으며 불필요한 영역 캡션이 다수 존재한다. 또한, 스토리 영상 콘텐츠인 영화 및 드라마 영상을 대상으로 한 연구가 매우 부족한 실정이다.

본 연구에서는 스토리 영상 콘텐츠 중 하나인 영화를 대상으로 영화 자동 이해를 위한 핵심 영역 중심의 이미지 캡셔닝 프레임워크를 제안한다. 제안 프레임워크는 DenseCap 모델을 기반으로 하며 캐릭터 식별 모듈, 핵심 영역 캡션 검출 알고리즘, 후처리기로 구성된다. 영화의 캐릭터 정보를

Collection @ daejin

핵심 영역 캡션 검출과 캡션 개선에 활용하기 위해 HoG 알고리즘과 DBSCAN을 사용하여 캐릭터 얼굴 학습 데이터 세트를 생성하고, EfficientNet 모델을 통해 영화 장면 속 캐릭터를 식별하는 캐릭터 식별 모듈을 설계 및 구현하였다. 또한, 영역 캡션 중요도에 영향을 미치는 변수로 '영역 박스 신뢰도 점수', '영역 면적', '영역과 장면 중심 간 거리', '객체 종류' 4가지를 고려하는 핵심 영역 캡션 검출 알고리즘을 제안한다. 이후, 후처리기에서 동일 객체에 대한 영역 캡션을 하나로 통합하고, 영역 캡션을 영화에 적합한 형태로 개선하는 작업 수행한다. 제안 프레임워크는 영화 장면의 중요한 영역 캡션을 효과적으로 검출하며, 기존 밀집 캡셔닝 모델보다 적은 수의 영역 캡션으로도 많은 정보를 제공하는 것을 확인하였다.

2. 논문의 구성

논문의 구성은 다음과 같다.

II. 관련 연구에서는 캐릭터 식별 모듈에서 사용되는 얼굴 인식 관련 모델과 이미지 캡셔닝 모델 중 제안 프레임워크의 기반이 되는 밀집 캡셔닝 모델에 대해 중점적으로 설명한다.

III. 제안 프레임워크에서는 본 연구에서 제안하는 프레임워크의 구성과 각 구성 요소에 대해 자세하게 설명한다.

IV. 실험 및 결과에서는 실험 환경과 실험 데이터에 대해 설명하고, 기존 밀집 캡셔닝 모델과 제안 프레임워크의 결과를 시각적으로 비교한다.

V. 결론에서는 실험 결과를 바탕으로, 본 연구가 갖는 의의와 한계점, 향후 연구에 대해 기술한다.

Collection @ daejin

Ⅱ. 관련 연구

1. 얼굴 인식

가. HOG(Histogram of Oriented Gradients)

HOG 특징은 이미지 내에서 물체를 검출하기 위한 컴퓨터의 영상 처리 알고리즘에 사용된다. 이 방법은 이미지의 국한된 부분 안에서 기울기 방향의 수를 계산하는 것이며, 밝기, 기울기, 에지 분포에 따라 그 값이 결정된다. 이러한 이미지 내부의 특징들은, 이미지를 셀(cell)이라는 서로 연결된 작은 영역으로 나누고 각 셀 내에 있는 픽셀(pixel)에 대한 기울기 방향, 에지 방향의 히스토그램을 만들어서 나타낸다. 이러한 히스토그램들이 모여서 이미지 내부의 특징을 나타낼 수 있다. 또한, 이미지 내부의 특징을 보다 잘 나타내기 위하여 블록(block)이라는 셀보다 큰 이미지 영역에 걸친 명암을 측정하여 블록 내의 모든 셀들을 정규화함으로써 조명이 변화하는 상황에서도 알고리즘이 일정한 성능을 유지할 수 있도록 한다[1, 8].

HOG 과정은 크게 3단계로 나뉜다. 첫 번째 단계는 1차원 중심점 이산 미분 마스크를 수평/수직 방향으로 적용하여 이미지 내부의 기울기 값을 계산한다[1, 8].

$$D_X = [-1 \quad 0 \quad 1], \qquad D_Y = [-1 \quad 0 \quad 1]$$

그리고 이미지 I가 주어졌을 때, 위의 식의 마스크와 이미지를 각각 x 방향과 y 방향으로 아래 식과 같이 convolution 연산한다[1, 11].

Collection @ daejin

$$I_X = I * D_X, \qquad I_Y = I * D_Y$$

위에서 구한 I_X와 I_Y를 이용하여 이미지의 기울기의 크기와 방향을 다음과 같이 계산한다[1, 11].

$$\text{기울기의 크기} : |G| = \sqrt{I_X^2 + I_Y^2}, \ \text{기울기의 방향} : \theta = \arctan\frac{I_Y}{I_X}$$

두 번째 단계에서는 각 셀의 히스토그램은 계산한다. 셀 내에 각 픽셀 값은 상기한 기울기 계산을 통해서 기울기의 방향이 연산이 되고 이 값들은 빈(bin) 수로 설정된 각각의 방향 히스토그램 대역들에 퍼지게 된다. 각 셀들은 이미지 내에서 사각형으로 이루어져 있고, 히스토그램 대역들은 0°에서 360° 또는 0°에서 180°까지 분포한다[1, 11].

세 번째 단계에서는 조명변화에 따라 검출 대상의 형상이 변하는 것에 대응하기 위하여 기울기의 크기를 국부적으로 정규화한다. 이 과정에서 셀들을 더 크고 공간적으로 연결된 블록들로 그룹화한다[1, 7]. 블록을 정규화하는 방법에는 크게 3가지가 있다. 주어진 블록 내의 모든 히스토그램을 포함하는 비정규화된 벡터를 ν, $\|\nu\|$를 k 놈(norm), e는 상수라고 했을 때 정규화 방법은 아래와 같다[1, 13].

$$L_1 - norm : f = \frac{\nu}{\|\nu\|_1 + e}$$

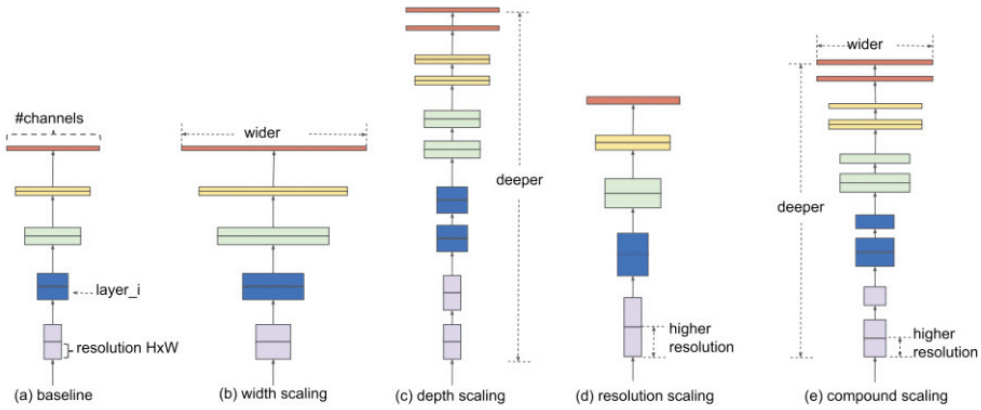
$$L_1 - \sqrt{} : f = \sqrt{\frac{\nu}{\|\nu\|_1 + e}}$$

$$L_2 - norm : f = \frac{\nu}{\|\nu\|_2^2 + e^2}$$

Collection @ daejin

나. EfficientNet

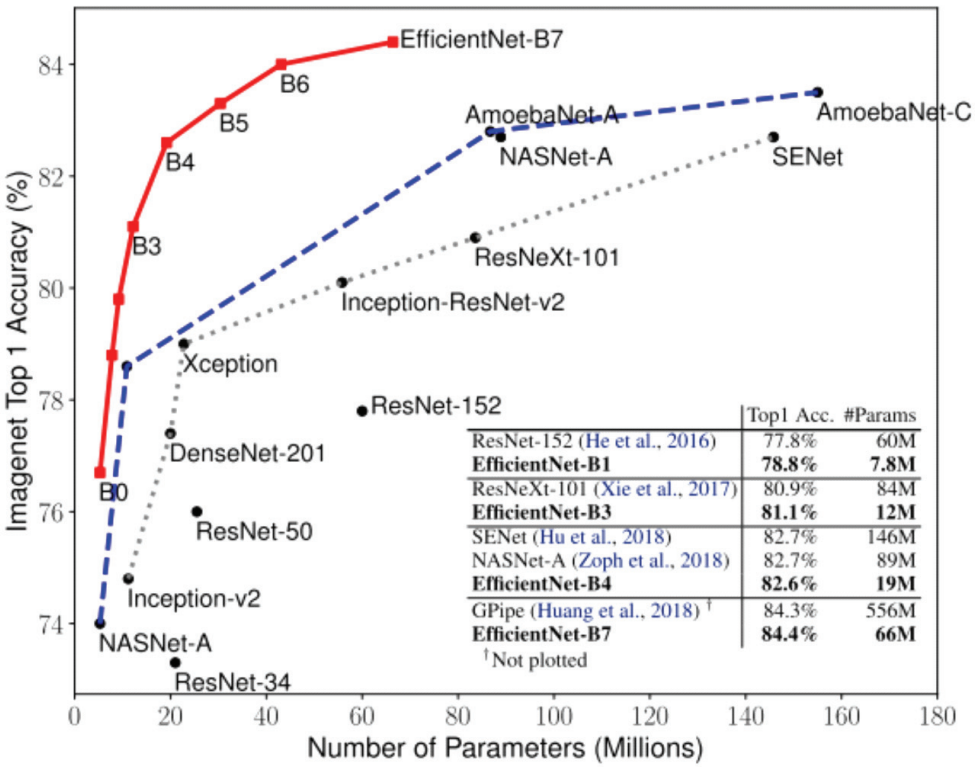
일반적으로, CNN(Convolutional Neural Network)의 정확도를 높이기 위해서 새로운 모델을 개발하는 방법이 있지만, 기존 모델을 바탕으로 복잡도(complexity)를 높이는 방법도 많이 사용한다. 기존 모델의 크기를 키우는 방법에는 아래의 [그림 2-1]과 같이 채널(channel)의 개수를 늘리는 width scaling과 층(layer)의 개수를 늘리는 depth scaling, 입력 이미지의 해상도를 높이는 resolution scaling 등이 있다. depth scaling을 통해 모델의 크기를 조절하는 대표적인 모델로는 ResNet이 있으며, width scaling을 통해 모델의 크기를 조절하는 대표적인 모델로는 MobileNet, ShuffleNet 등이 있다. 그러나 3가지 scaling을 모두 적용하는 경우는 거의 없다.



[그림 2-1] 모델 크기 조정[12]

Tan et al.(2019)은 모델을 고정하고 depth, width, resolution 3가지를 동시에 고려하는 compound scaling을 제시하였으며, MnasNet과 유사한 탐색 공간(search space)에서 AutoML을 통해 모델을 탐색하고, 이 과정에서 찾은 작은 모델을 EfficientNet이라고 하였다. AutoML을 통해 찾은 EfficientNet 모델들과 기존의 모델들을 비교한 결과는 아래의 [그림 2-2]과 같다.

Collection @ daejin



[그림 2-2] 기존 CNN 모델과 EfficientNet 모델의 정확도 및 파라미터 수[12]

EfficientNet은 상대적으로 적은 파라미터(parameter) 수와 연산량으로 기존의 CNN 모델들을 상회하는 정확도를 보이며, ImageNet 데이터 세트에서 가장 높은 정확도를 보였던 GPipe 보다 더 높은 정확도를 보임을 확인하였다.

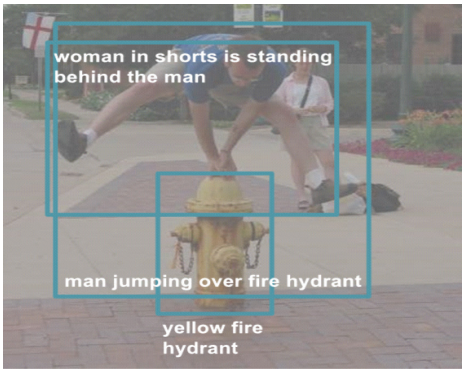
2. 이미지 캡셔닝(Image Captioning)

이미지 캡셔닝은 컴퓨터 비전에서 핵심 연구 과제 중 하나로 많은 관련 연구가 존재한다. 본 장에서는 제안 프레임워크의 기반이 되는 밀집 캡셔닝의 대표적인 모델들과 학습 데이터 세트를 설명한다.

Collection @ daejin

가. 비주얼 게놈(Visual Genome)

이미지 캡셔닝을 위한 기존의 데이터 세트는 전체 이미지에 대한 캡션(caption)을 제공하거나[3, 15] 이미지 내 영역(region)에 대한 레이블(label)은 제공하지만 완전한 문장 형태의 캡션은 제공하지 않는다[10]. 비주얼 게놈(Visual Genome)의 영역 캡션(region caption) 데이터 세트는 MS COCO와 YFCC100M 데이터 세트에서 이미지를 수집하였으며, 크라우드 소싱(crowd sourcing) 플랫폼인 Amazon Mechanical Turk를 통해 분산 인력이 이미지에 영역 경계 박스(bounding box)를 생성하고 해당 영역에 대한 묘사를 자연어(natural language) 형태의 캡션으로 작성하도록 하여 구축되었다. 비주얼 게놈의 영역 캡션 데이터 세트의 예는 [그림 2-3]와 같으며, 현재 108,077개의 이미지와 약 5,400,000개의 영역 캡션을 포함한다[6].



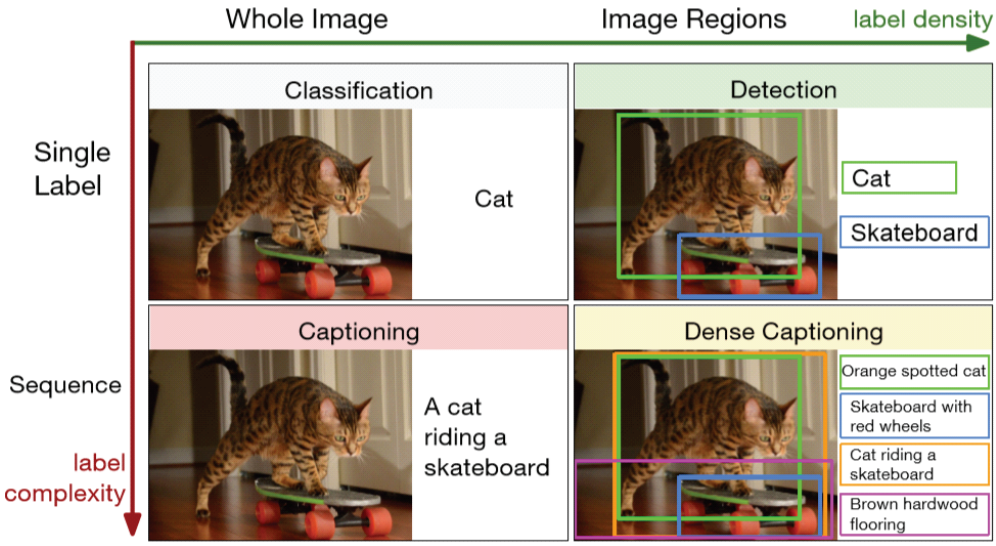
[그림 2-3] 비주얼 게놈 영역 캡션 데이터 예시[6]

나. DenseCap

이미지에 대한 묘사를 자연어 형태로 생성하는 과제는 크게 두 가지 방향으로 연구되었다. 이미지의 중요한 여러 영역을 효율적으로 식별하고 레이블을 생성하는 객체 검출(object detection)과 레이블의 복잡도를 확대

Collection @ daejin

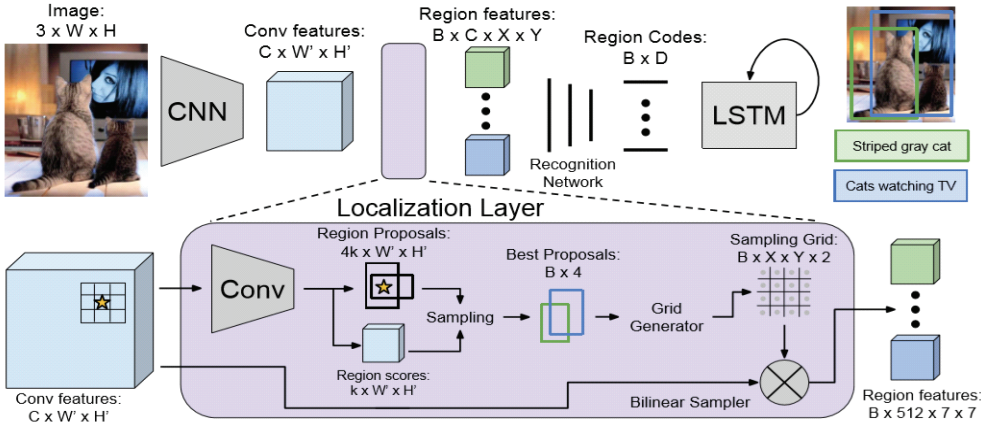
하여 전체 이미지에 대한 묘사를 단어 시퀀스(sequence of words) 형태로 생성하는 이미지 캡셔닝이 있다. 그러나, 두 연구는 아래의 [그림 2-4]과 같이 레이블의 밀도(density)와 복잡도(complexity)를 두 축으로 서로 독립적으로 진행되었다.



[그림 2-4] 밀집 캡셔닝[4]

DenseCap[4]은 이러한 객체 검출과 이미지 캡셔닝 모델을 하나의 공동 프레임워크로 통합하는 모델이다. DenseCap은 밀도 캡셔닝(dense captioning)을 위한 Fully Convolutional Localization Network(FCLN)를 제안하였다. CNN과 RNN(Recurrent Neural Network)으로 구성된다는 점에서 기존 이미지 캡셔닝의 인코더-디코더(Encoder-Decoder) 방식을 따른다. 그러나 지역화 계층(Localization layer)을 추가하여 영역 수준의 학습과 예측을 가능하게 한다. 전체적인 DenseCap의 개요는 아래의 [그림 2-5]와 같다.

Collection @ daejin

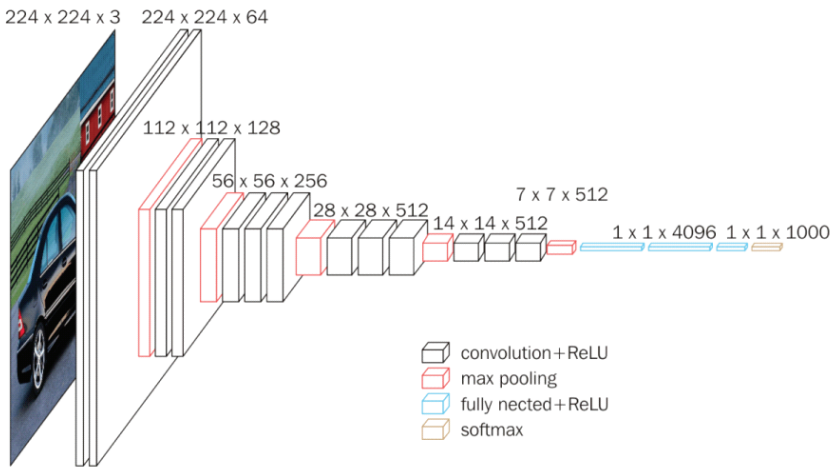


[그림 2-5] DenseCap 개요[4]

DenseCap의 CNN은 [그림 2-6]의 VGG-16[11]을 사용한다. 5개의 2×2 최대 풀링 계층(max pooling layer)와 13개의 3×3 컨볼루션 계층(convolution layer)으로 구성되지만, 최종 풀링 계층을 제거하여 $3\times W\times H$ 형태의 입력 이미지는 $C=512$, $W'=\left\lfloor \frac{W}{16} \right\rfloor$, $H'=\left\lfloor \frac{H}{16} \right\rfloor$ 인 $C\times W'\times H'$ 형태의 텐서(tensor)로 변환된다. 이러한 텐서는 지역화 층의 입력이 된다.

Faster R-CNN 기반의 지역화 계층은 RoI(Regions of Interest) 풀링 메커니즘을 양선형 보간법(bilinear interpolation)으로 대체하여 예측된 영역 좌표(region coordinates)를 통해 역전파(backpropagation)를 할 수 있게 하였다. 지역화 계층은 $C\times W'\times H'$ 형태의 텐서를 입력으로 받아, B개의 관심 영역(RoI, Regions of Interest)을 추출한다. 이후, $B\times4$ 행렬의 영역 좌표(region coordinate)와 B 길이의 벡터 형태인 영역 점수(region score), $B\times C\times X\times Y$ 형태 영역 특징(region feature) 텐서를 반환한다.

Collection @ daejin



[그림 2-6] VGG-16 구성도[11]

RNN 언어 모델(language model)로는 LSTM(Long Short Term Memory)을 채택하고, 영역 코드(region code)가 사용된다. 예를 들어, 토큰(token) $S_1,...,S_T$의 학습 시퀀스(sequence)가 주어지면, RNN 모델에 $T+2$ 길이의 워드(word) 벡터 $x_{-1}, x_0, x_1,..., x_T$를 제공한다. 이 때, $x_{-1} = CNN(I)$는 선형 계층(linear layer)으로 인코딩되고 비선형 ReLU(Rectified Linear Unit)를 통해 추출된 영역 코드이다. x_0는 특수 START 토큰에 해당하고, x_t는 각 토큰 $S_t, t = 1,..., T$를 인코딩한다. 아래의 식을 사용해서 은닉 상태(hidden states) 시퀀스 h_t와 출력 벡터 y_t를 계산한다.

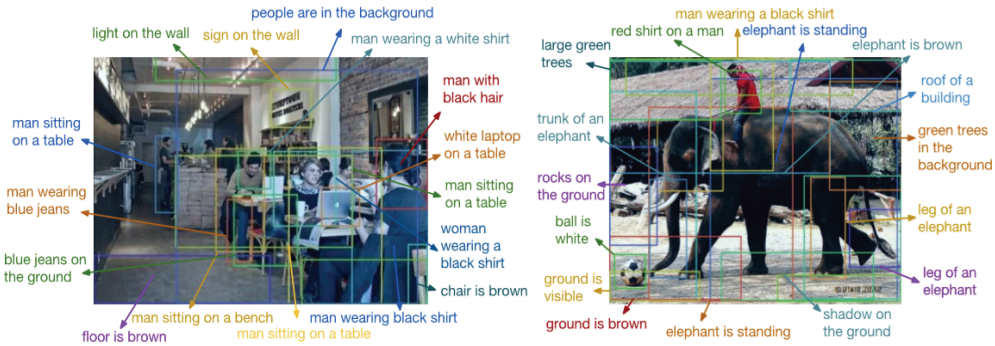
$$h_t, y_t = f(h_{t-1}, x_t)$$

벡터 y_t의 크기는 $|V|+1$이며, 여기서 V는 토큰 어휘이고 추가 토큰은 특수 END 토큰을 나타낸다.

테스트 타임(test time)에 시각적 정보인 x_{-1}을 RNN에 제공한다. 각 타임 스텝(time step)에서 가장 가능성이 높은 토큰을 샘플링하고, 다음 타임

Collection @ daejin

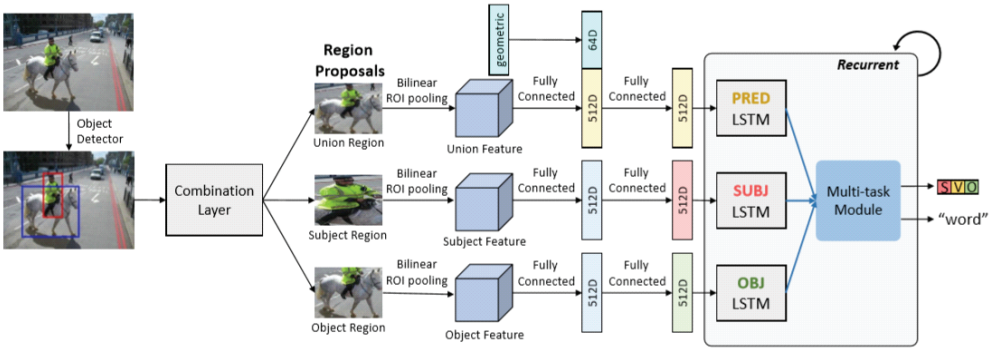
스텝에서 RNN에 공급하여 특수 END 토큰이 샘플링될 때까지 프로세스를 반복한다.



[그림 2-7] DenseCap 모델 생성 캡션 샘플[4]

다. Dense Relational Captioning

Dense Relational Captioning[5]은 이미지 내 객체들의 관계 기반의 캡션을 생성하는 모델이다. 영어 단어에 주어(subject), 술어(predicate), 목적어(object) 범주의 POS(Part Of Speech) 태그(tag)를 할당하고, 각 태그에 대해 3개의 순환 유닛(recurrent unit)으로 구성하여 태그 예측과 캡션 생성을 공동으로 수행하는 MTTSNet(Multi-Task Triple-Stream Network)을 제안하였다.



[그림 2-8] Dense Relational Captioning 구성도[5]

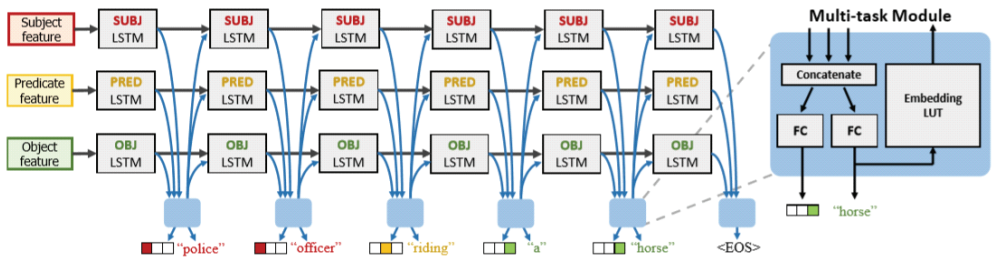
Collection @ daejin

Dense Relational Captioning 모델의 인코더는 DenseCap 모델과 동일한 구조를 채택한다. 반면, DenseCap 모델과의 차이점은 관계 제안(relational proposal)을 생성하기 위해 추출한 B개의 관심 영역들의 조합을 만들어 B(B+1)개의 영역 쌍을 생성한다. 이러한 역할을 하는 층을 조합 계층(conbination layer)으로 부른다. 또한, Yang et al.(2017)의 방법을 채택하여 주체(subject)와 객체(object)를 결합(union)한 영역인 bu(box union)를 추가적으로 활용한다. 또한, 상대적인 공간 정보를 제공하기 위해 주체와 객체 박스 쌍에 대한 기하학적 특징(geometric feature)을 완전 연결 계층(fully connected layer) 이전의 결합 특징(union feature)에 추가한다. 두 개의 경계 박스 bs(box subject)와 bo(box object)가 주어지면 기하학적 특징 r은 Peyre et al.(2017)과 유사하게 아래의 식으로 정의된다. 주체, 객체, 결합 영역에서 추출한 3개의 특징은 각 LSTM에 제공된다.

$$r = \left[\frac{x_o - x_s}{\sqrt{w_s h_s}}, \frac{y_o - y_s}{\sqrt{w_s h_s}}, \sqrt{\frac{w_o h_o}{w_s w_s}}, \frac{w_s}{h_s}, \frac{w_o}{h_o}, \frac{b_s \cap b_o}{b_s \cup b_o} \right] \in R^6$$

Dense Relational Captioning 모델은 RNN 언어 모델로서 관계 캡션(relational caption)을 생성하는 모델인 MTTSNet를 제안하였다. 주체와 객체의 영역 코드는 캡션에 있는 주어와 목적어에 관한 단어와 밀접하게 연관되어 있는 반면, 결합과 기하학적 특징은 술어 예측에 기여할 수 있다. 제안된 triple-stream LSTM 모듈은 각 주체, 객체, 결합 영역 코드를 담당하는 세 개의 LSTM으로 구성된다. 세 개의 LSTM은 각각 하나의 처리된 표현(representation)을 생성하며, 세 개의 표현을 통합하여 하나의 단어를 예측한다. 이러한 구조는 단일 LSTM보다 더 많은 유연성을 허용하며, 주어, 술어, 목적어 태그 순으로 관계 캡션 단어 시퀀스를 생성하도록 한다.

Collection @ daejin

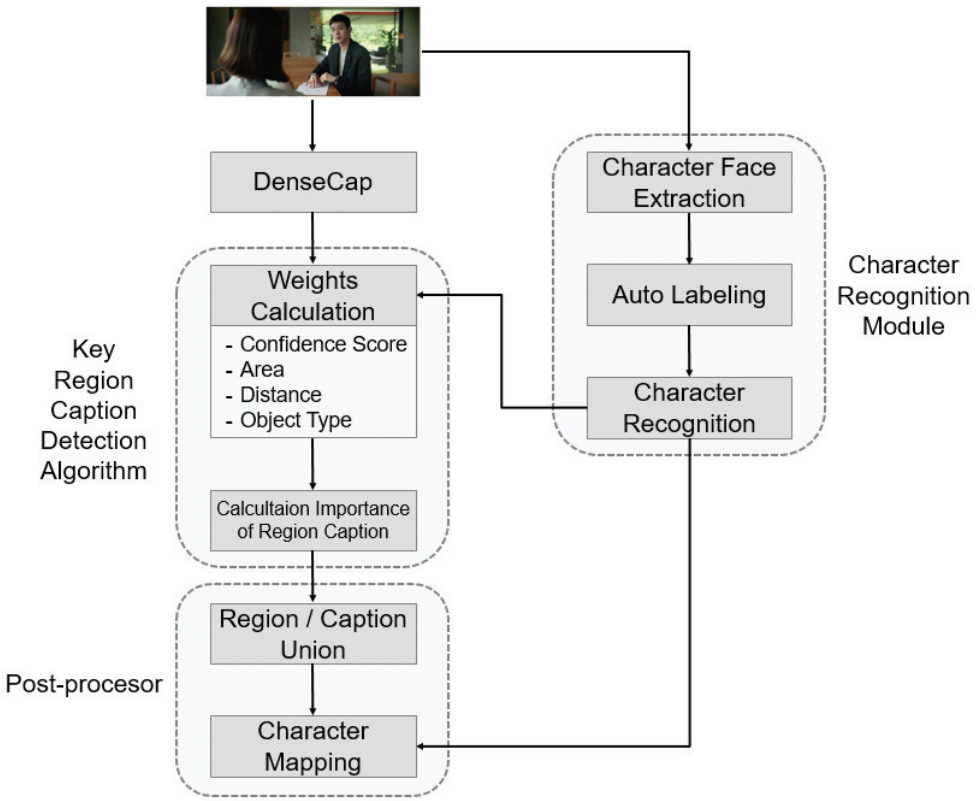


[그림 2-9] triple-stream LSTM 구조[5]

Collection @ daejin

Ⅲ. 제안 프레임워크

1. 프레임워크 구성

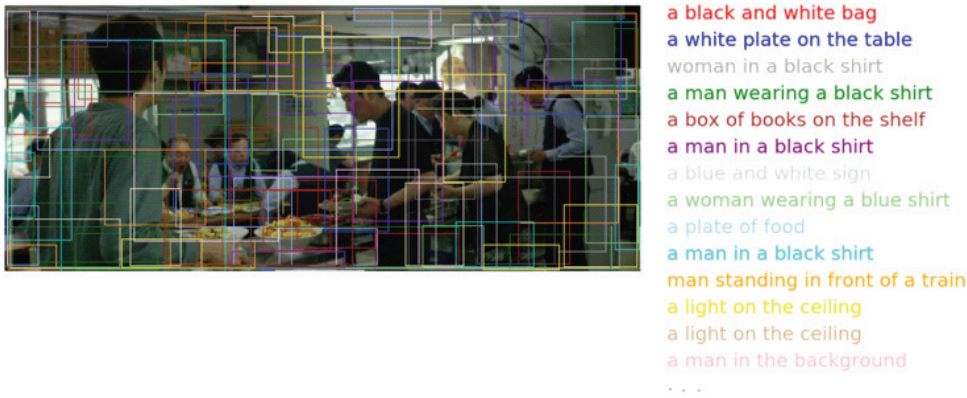


[그림 3-1] 제안 프레임워크 수행 절차

본 연구에서 제안하는 프레임워크의 수행 절차는 [그림 3-1]과 같다. 먼저, 영화 영상에서 1초당 1장의 프레임을 추출하여 프레임 단위의 장면(scene) 이미지 집합을 생성한다. 장면 이미지는 이미지 캡셔닝 대상이 되며, 이미지를 너비(width)를 720픽셀(pixel)로 변환하는 전처리 과정을 수행한다. 본 연구에서는 DenseCap 모델을 각 장면 이미지에 대해 1,000개의

Collection @ daejin

영역 제안(region proposal)을 생성하고, 최초 NMS(Non Maximum Supression)의 임계값(threshhold)을 0.7, 최종 NMS 임계값을 0.3으로 설정하여 영역의 불필요한 중복을 줄였다. [그림 3-2]과 같이 한 장면 이미지에 대해 평균 105개의 영역 캡션을 생성한다.



[그림 3-2] 영화 장면 이미지에 대한 DenseCap 영역 캡션

본 연구에서는 장면 이미지 내에서 등장하는 캐릭터의 정보를 영역 캡션 중요도에 반영하기 위한 캐릭터 식별 모듈(Character Recognition Module)을 설계 및 구현하였다. 캐릭터 식별 모듈은 3.2절에서 자세하게 설명한다. 또한, 본 논문에서 제안하는 핵심 영역 캡션 검출 알고리즘(Key Region Caption Finding Algorithm)은 영역 캡션의 중요도에 영향을 미치는 변수로 '영역 박스 신뢰도 점수', '영역 면적', '영역과 장면 중심 간 거리', '객체 종류' 4가지를 고려하여 다양한 영역 캡션 중 상대적으로 중요한 핵심 영역 캡션을 찾는다. 이러한 핵심 영역 검출 알고리즘은 3.3절에서 자세하게 다룬다. 마지막으로 핵심 영역 및 캡션을 개선하고 영화에 적합한 캡션으로 변환하는 작업을 후처리기(Post-processor)에서 수행하며, 후처리기에 대한 내용은 3.4절에서 설명한다.

Collection @ daejin

2. 캐릭터 식별 모듈(Character Recognition Module)

본 연구에서 제안하는 캐릭터 식별 모듈은 얼굴 인식 모델을 사용하여 장면 이미지에 등장하는 캐릭터의 이름과 얼굴 박스를 예측한다. 이러한 캐릭터 식별 모듈을 통해서 예측된 캐릭터의 이름과 얼굴 박스는 3.3절에서 '객체 종류'가 인물인 경우 영화 캐릭터에 따른 중요도를 부여하기 위해 사용되며, 3.4절 후처리기에서 영역 캡션의 사람을 지칭하는 단어를 캐릭터 이름으로 변환할 때 사용한다.

얼굴 인식 모델의 학습 데이터 세트를 생성하기 위해 먼저 장면 이미지 집합 얼굴 이미지를 수집하는 과정을 수행한다. 특징 추출 알고리즘인 Haar-Cascade와 HOG로 장면 이미지 집합에서 인식한 얼굴을 크롭(crop)하여 저장하였다. 아래의 <표3-1>와 같이 Haar-Cascade 알고리즘은 총 6,211장의 이미지를 수집하였으며, 정확하게 인식한 얼굴 이미지는 2,384장으로 38.38%의 정확도를 보였다. 반면, HOG 알고리즘은 총 2,327장의 이미지를 수집하였으며, 정확하게 인식한 얼굴 이미지는 2,299장으로 98.80%의 정확도를 보이며, Haar-Cascade 알고리즘보다 정확도가 60.42% 높았다. 본 연구에서는 얼굴 이미지를 수집하기 위해 HOG 알고리즘을 채택하여 사용한다.

<표 3-1> Haar-Cascade와 HOG 알고리즘 얼굴 인식 정확도

	Haar-Cascade	HOG
총 이미지 수	6,211장	2,327장
얼굴 이미지 수	2,384장	2,299장
정확도	38.38%	98.80%

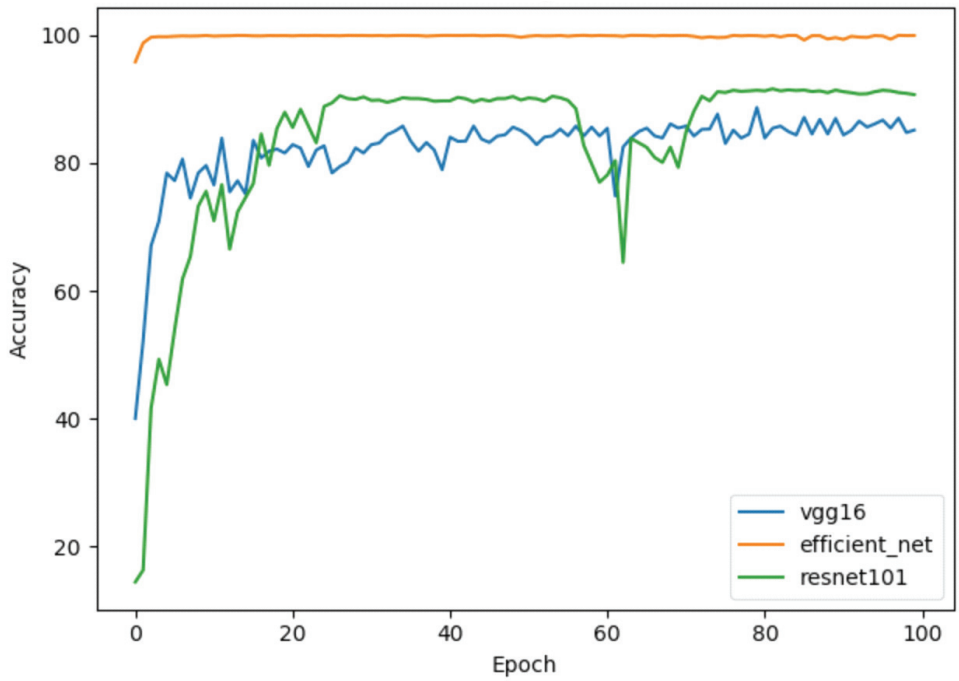
Collection @ daejin

[그림 3-3] Haar-Cascade 알고리즘과 HOG 알고리즘의 얼굴 인식 결과

수집한 얼굴 이미지는 지도 학습(supervised learning)에 필요한 레이블이 존재하지 않기 때문에 레이블을 부여하는 과정이 수행되어야 한다. 레이블을 부여하기 위해 얼굴 이미지의 HOG 특징 벡터를 밀도 방식 클러스터링(clustering) 기법 중 하나인 DBSCAN을 수행하여 유사한 얼굴 이미지를 군집한다. 군집들이 나타내는 캐릭터의 이름을 레이블로 부여한다. 이후, 엑스트라의 얼굴 이미지를 제외하고 잘못 군집된 얼굴 이미지를 직접 재분류하는 작업을 진행한다. 마지막으로 얼굴 이미지를 자르기(crop), 회전(rotate), 뒤집기(flip), 밀기(translate), 크기 조정(resize) 등으로 변형하는 어그멘테이션(agumentation)을 통해 1장당 30장의 이미지를 확보한다.

본 연구에서는 앞서 생성한 얼굴 이미지 학습 데이터 세트에 대한 CNN 모델인 VGG-16, ResNet101, EfficientNet-B5 3가지 모델의 정확도 성능을 비교하였다. 동일한 실험 환경과 조건에서 학습을 진행한 결과 아래 [그림 3-4]와 같이 EfficientNet-B5 모델이 99.85%의 정확도로 가장 높았으며, 다음 ResNet101 모델이 91.612%, VGG-16 모델이 85.15%의 정확도를 보였다. 따라서 EfficientNet을 얼굴 인식 모델로 채택하여 사용한다. 캐릭터 식별 모듈의 수행 절차는 [그림 3-5]와 같다.

Collection @ daejin



[그림 3-4] CNN 모델들의 얼굴 인식 정확도

[HoG 알고리즘 결과] [DBSCAN 결과] [EfficientNet 예측 결과]

[그림 3-5] 캐릭터 식별 모듈의 수행 절차

3. 핵심 영역 캡션 검출 알고리즘(Key Region Caption Detection Algorithm)

본 장에서는 DenseCap을 통해 생성된 영역 캡션들의 중요도를 계산하고 중요도를 바탕으로 핵심 영역 캡션을 추출하는 알고리즘에 대한 내용을

Collection @ daejin

다룬다. 영역 캡션의 중요도에 영향을 미치는 변수들와 중요도 계산 방법에 대한 세부 내용을 아래의 절에서 자세히 설명한다.

가. 영역 박스 신뢰도 점수

영역 박스의 신뢰도 점수(confidence score)는 영역 박스 안에 객체가 존재할 확률이 높을수록, 영역 박스와 정답 박스(ground truth box)가 일치할수록 큰 값을 갖는 영역 박스에 대한 신뢰도 지표이다. 따라서, 영역 박스의 신뢰도 점수가 높을수록 영역 박스가 객체를 정확하게 포함하고 있을 확률이 높으므로 영역 박스의 신뢰도 점수가 클수록 중요한 영역이라고 가정한다. 영역 박스의 신뢰도 점수는 아래의 식으로 계산한다. $Confidence\,Score_i$는 i번째 영역의 신뢰도 점수를 의미하며, $\Pr(Object)$는 영역 박스 안에 객체가 존재할 확률을 나타낸다. $IoU(Truth, B_i)$는 정답 박스 $Truth$와 i번째 영역 박스 B_i가 겹치는 영역을 비율로 계산한 것이다.

$$Confidence\,Score_i = \Pr(Object) \times IoU(Truth, B_i)$$

[그림 3-6]은 영역 박스 신뢰도 점수만을 고려한 상위 5개 영역 캡션을 보여준다. 첫 번째 장면 이미지에서 추출된 영역 캡션은 ‘a woman in a bathroom’, ‘a towel hanging on the wall’, ‘man wearing black shirt’, ‘woman wearing black pants’, ‘a white toilet bowl’로 영역 박스가 ‘woman’, ‘man’, ‘towel’, ‘toilet bowl’ 등 장면에서 등장하는 다양한 객체를 정확하게 포함한다. 두 번째 장면 이미지에서 추출된 영역 캡션은 ‘a man wearing a blue shirt’, ‘a large green tree’, ‘a cloudy blue sky’, ‘man flying a kite’, ‘a building with a red roof’로 영역 박스가 장면에서 등장하는 주요 객체인 ‘man’, ‘tree’, ‘sky’, ‘kite’, ‘building’을 정확하게 포함하는 것을 확인 할 수 있다.

Collection @ daejin

[그림 3-6] 영역 박스 신뢰도 점수를 고려한 상위 5개 영역 캡션

Collection @ daejin

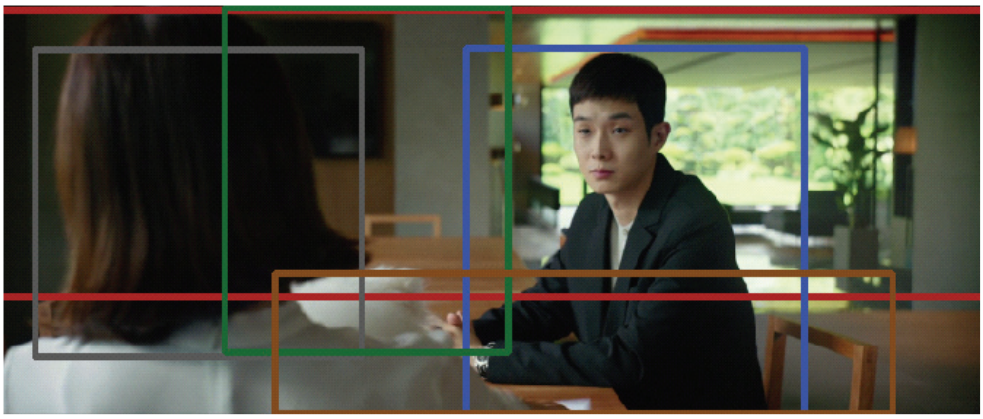
나. 영역 면적

일반적으로 영화에서는 장면의 중요한 객체를 강조하기 위해 클로즈업 등의 촬영 기법으로 전체 장면에서 차지는 면적을 크게 한다. 영역의 면적이 클수록 영역이 나타내는 객체가 장면에서 중요한 객체일 가능성이 크다는 가정이다. 영역의 면적은 아래의 식으로 계산한다. 여기서, $Area_i$는 i번째 영역의 면적을 의미하며, $B_{i,W}$와 $B_{i,H}$는 i번째 영역 박스의 너비와 높이를 의미한다.

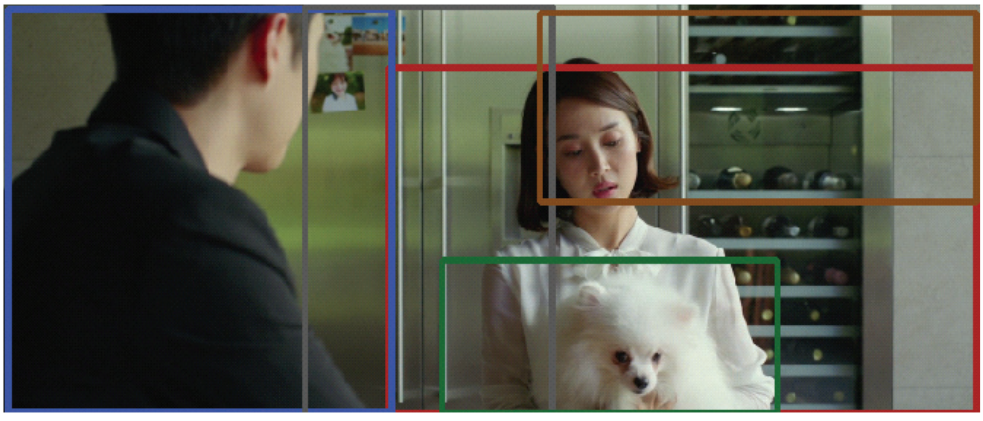
$$Area_i = B_{i,W} \times B_{i,H}$$

[그림 3-7]은 영역 면적만을 고려한 상위 5개 영역 캡션을 보여준다. 첫 번째 장면 이미지에서 추출된 영역 캡션은 ‘people standing in front of a window’, ‘man in a suit’, ‘woman has long hair’, ‘a large door’, ‘a man sitting on a chair’로 장면에서 차지는 면적이 큰 ‘people’, ‘window’, ‘man’, ‘woman’, ‘door’, ‘chair’와 같은 객체들을 설명하는 캡션이 나타났으며, 두 번째 장면 이미지에서 추출된 영역 캡션은 ‘the door is open’, ‘man wearing black shirt’, ‘a door in the room’, ‘white dog with a collar’, ‘a white and black oven’으로 장면에서 대부분 면적을 차지하고 있는 ‘door’, ‘man’, ‘dog’, ‘oven’과 같은 객체들을 설명하는 캡션들이 나타났다.

Collection @ daejin



1. people standing in front of a window
2. man in a suit
3. woman has long hair
4. a large door
5. a man sitting on a chair



1. the door is open
2. man wearing black shirt
3. a door in the room
4. white dog with a collar
5. a white and black oven

[그림 3-7] 영역 면적을 고려한 상위 5개 영역 캡션

Collection @ daejin

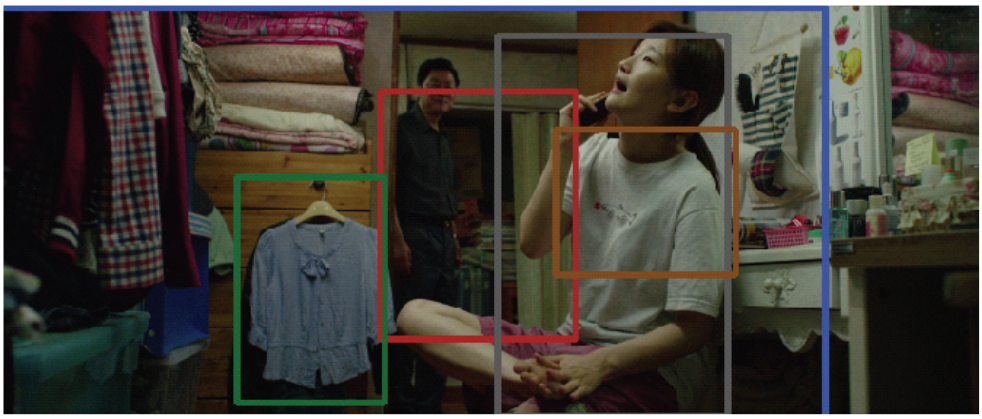
다. 영역과 장면 중심 간 거리

또한, 장면의 중요한 객체일수록 카메라의 중앙에 위치하도록 촬영하는 경우가 일반적이다. 따라서 객체가 장면의 중심과 가까울수록 중요한 객체일 가능성이 크다고 가정한다. 본 연구에서는 영역과 장면의 중심 간의 거리를 아래의 식과 같이 유클리드 거리 측정 방법을 사용하며 계산하였으며, 역수를 취하여 거리가 짧을수록 큰 값을 갖도록 하였다. 여기서 $S_{centerX}$와 $S_{centerY}$는 장면의 중심 X 좌표와 Y 좌표를 의미하며, $B_{i,centerX}$와 $B_{i,centerY}$는 i번째 영역 박스의 중심 X 좌표와 Y 좌표를 의미한다. 최종적으로 구해진 i번째 영역 거리를 $Distance_i$라고 한다.

$$Distance_i = \frac{1}{\sqrt{(S_{centerX} - B_{i,centerX})^2 + (S_{centerY} - B_{i,centerY})^2}}$$

[그림 3-8]은 영역과 장면 중심 간 거리가 가까운 상위 5개 영역 캡션을 보여준다. 첫 번째 장면 이미지에서는 ‘man wearing a blue shirt’, ‘two people standing in a room’, ‘woman in a green shirt’, ‘a blue and white jacket’, ‘white shirt on the woman’과 같은 영역 캡션이 추출되었으며 해당 장면에서 중요한 객체인 ‘man’, ‘people’, ‘woman’이 장면 중심과 가까워 상위 영역 캡션에 나타났다. 또한, ‘jacket’, ‘shirt’와 같은 객체도 장면 중심과 가깝기 때문에 상위에 등장했다. 두 번째 장면 이미지에서는 ‘a man and a woman’, ‘a box of paper’, ‘man with dark hair and a shirt’, ‘a box of cardboard box’와 영역 캡션이 추출되었으며 해당 장면에서 중요한 객체인 ‘man’, ‘woman’, ‘box’가 상위에 나타났으며, 주변에 있는 인물들은 상대적으로 장면의 중심과 거리가 멀어 상위 영역 캡션에 등장하지 않았다.

Collection @ daejin



1. man wearing a blue shirt
2. two people standing in a room
3. woman in a green shirt
4. a blue and white jacket
5. white shirt on the woman

1. a man and a woman
2. a box of paper
3. a box of paper
4. man with dark hair and a shirt
5. a box of cardboard box

[그림 3-8] 영역과 장면 중심 간 거리를 고려한 상위 5개 영역 캡션

Collection @ daejin

라. 객체 종류

본 논문에서는 캡션에서 추출한 명사를 객체라고 정의한다. 객체는 인물과 사물, 배경으로 구분되며, 인물의 경우 해당 인물의 비중에 따라 주연, 조연, 엑스트라로 구분할 수 있다. 객체의 종류에 따라 중요도가 다르다고 가정하고, 본 연구에서는 아래의 <표 3-2>와 같이 객체 종류에 따른 중요도를 정의하였다.

<표 3-2> 객체 종류에 따른 중요도

	주연	조연	엑스트라	사물	배경
중요도	3	2	1	1	0

본 연구에서는 [그림 3-9]와 같이 3.2절에서 예측한 캐릭터의 얼굴 박스가 영역 박스에 70% 이상 포함되고, 영역이 나타내는 객체의 종류가 인물이면 객체를 예측한 캐릭터로 판단하고 해당 캐릭터에 부합하는 중요도를 부여한다. 객체의 종류가 인물이지만 아래와 같은 경우에는 객체의 종류를 엑스트라로 구분한다. 또한, 객체의 종류가 인물이 아닌 경우 사물 또는 배경으로 구분한다.

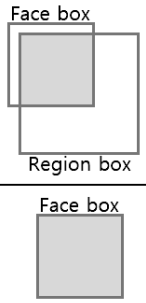
1) 인물의 얼굴이 명확하지 않은 경우

2) 얼굴 인식 모델로 식별되지 않는 인물

3) 얼굴 박스가 영역 박스에 포함되는 비율이 70% 미만인 경우

Collection @ daejin

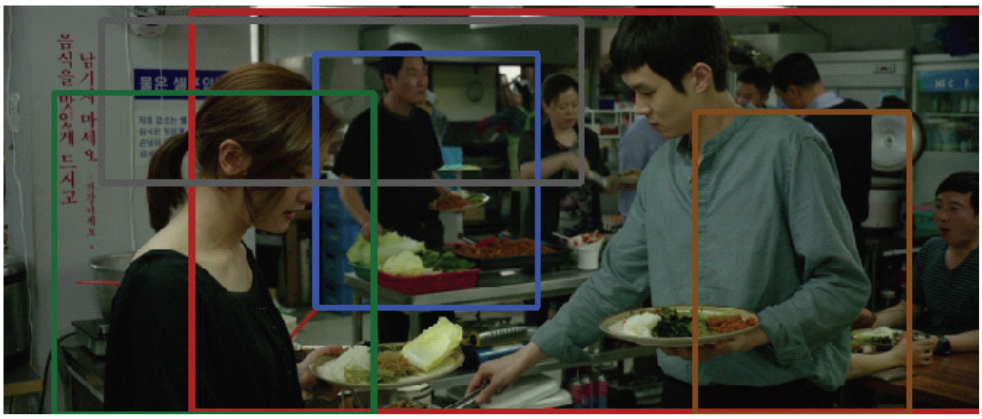
$$Intersection\ of\ Face = \frac{Area\ of\ Overlap}{Area\ of\ Face}$$



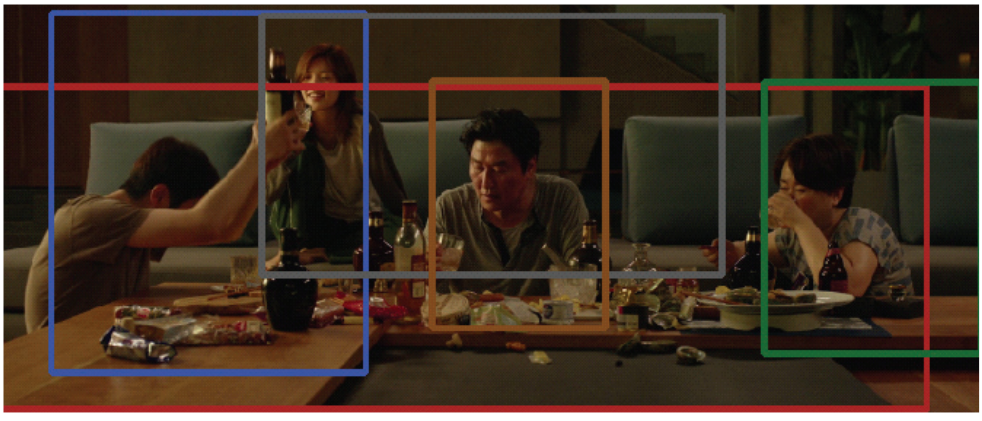
[그림 3-9] 얼굴 박스와 영역 박스의 교차 면적 비율

아래 [그림 3-10]은 객체 종류에 따른 중요도가 높은 상위 5개 영역 캡션을 보여준다. 첫 번째 장면 이미지에서는 ‘people at a table’, ‘woman in a black shirt’, ‘a woman with a red hair’, ‘woman with long hair’, ‘man wearing blue shirt’ 영역 캡션이 추출되었다. 장면에 다양한 객체들이 등장하지만 사물이나 배경보다 ‘people’, ‘woman’, ‘man’과 같은 인물이 우선적으로 나타났으며, 중요도가 높은 인물일수록 상위에 위치한다. 두 번째 장면 이미지에서는 ‘people sitting on a table’, ‘a woman sitting on a couch’, ‘a woman in a blue shirt’, ‘woman sitting on a couch’, ‘man in a blue shirt’가 핵심 영역 캡션이 추출되었으며 마찬가지로 다양한 객체 중 ‘people’, ‘woman’, ‘man’ 등 인물에 대한 영역 캡션이 상위 영역 캡션으로 나타난 것을 확인할 수 있었다.

Collection @ daejin



1. people at a table
2. woman in a black shirt
3. a woman with a red hair
4. woman with long hair
5. man wearing blue shirt



1. people sitting on a table
2. a woman sitting on a couch
3. a woman in a blue shirt
4. woman sitting on a couch
5. man in a blue shirt

[그림 3-10] 객체 종류를 고려한 상위 5개 영역 캡션

Collection @ daejin

마. 중요도 계산

앞에서 계산한 '영역 박스 신뢰도 점수', '영역 면적', '영역과 장면 중심 간 거리', '객체 종류'는 서로 단위가 다른 변수이기 때문에 영역 캡션 중요도에 미치는 영향의 크기 다르다. 따라서 아래의 식과 같이 각 변수에 대해 표준화 스케일링(standard scaling)을 적용하여 표준 정규 분포로 변환한다.

$$Standard\ Scaling(X) = \frac{X-\mu}{\sigma}$$

표준화 스케일링을 거친 각 변수를 $Confidence\ Score'$, $Area'$, $Distance'$, $Object\ Type'$으로 표기한다.

$$Confidence\ Score' = Standard\ Scaling(Confidence\ Score)$$
$$Area' = Standard\ Scaling(Area)$$
$$Distance' = Standard\ Scaling(Distance)$$
$$Object\ Type' = Standard\ Scaling(Object\ Type)$$

i번째 영역 캡션의 중요도는 아래의 식과 같이 스케일링을 거친 i번째 영역의 변수들의 합으로 계산한다.

$$Importance_i = Confidence\ Score'_i + Area'_i + Distance'_i + Object\ Type'_i$$

Collection @ daejin

4. 후처리기(Post-processor)

후처리 단계에서는 제안 프레임워크를 통해 추출된 상위 영역과 캡션의 개선을 위해 동일 객체를 나타내는 영역과 캡션을 통합하고, 영역 캡션 내 인물을 지칭하는 단어를 캐릭터 이름으로 변환하는 등의 작업을 수행한다. 후처리는 아래와 같은 방법으로 진행된다.

1) 두 영역이 동일 객체를 나타내고 IoU가 0.2 이상일 경우, 두 영역이 동일 객체에 대한 영역 캡션으로 판단하고 영역 박스와 캡션을 통합한다.

2) 영역이 나타내는 객체가 인물이고 캐릭터 식별 모듈이 예측한 캐릭터의 얼굴 박스가 영역 박스에 70% 이상 포함되는 경우, 아래의 <표 3-3>와 같은 방법으로 캡션에서 인물을 나타내는 단어를 캐릭터 이름으로 변환한다.

3) 캐릭터 이름 앞에 붙은 a와 the와 같은 관사는 제거한다.

<표 3-3> 영역 캡션 캐릭터 변환 방법

	영역 캡션이 나타내는 인물 수		
영역 박스 내 예측된 캐릭터	1명	2명	나머지
A	A	A and the other	A and the other
A, B	A or B	A and B	A and B and the other
3명 이상	A or B or …	A or B or …	A and B and … and the other

Collection @ daejin

IV. 실험 및 결과

1. 실험 환경 및 데이터

실험을 위해 <표 4-1>, <표 4-2>와 같이 하드웨어 및 소프트웨어 환경을 구성하였다. CPU는 Intel(R) Xeon(R) Gold 5120을 사용하였으며 VGA는 Tesla V100 SXM2 32GB 2대로 구성하였다. 또한, 운영체제는 Ubuntu 16.04 LTS를 사용했으며, CUDA 10.1와 cuDNN 7.5.1을 설치하였다. 제안 프레임워크는 Lua와 Python을 사용하여 Torch7와 Tensorflow 2.1 환경에서 구현되었다.

본 연구에서는 봉준호 감독의 '기생충' 영화를 분석 대상으로 선정하여 실험을 진행하였다. 분석에 사용한 영화는 1,920x804의 영상이며, 1초당 1장의 프레임을 추출하여 총 7,922장의 장면 이미지를 수집하였다. 이후, 1920x804의 이미지를 720x301 크기로 조정하는 전처리 과정을 수행하였다.

캐릭터 식별 모듈의 식별 대상이 캐릭터는 '기생충' 영화의 등장인물 메타데이터를 활용하여 아래의 <표 4-3>과 같이 주연 6명, 조연 4명 총 10명으로 정의하였다. 얼굴 인식을 위한 모델의 학습 데이터 세트를 생성하기 위해 장면 이미지 집합에서 HOG 알고리즘으로 인식한 얼굴을 크롭하여 수집한 2,327장 중 잘못 인식한 28장을 제외하여 총 2,299장의 얼굴 이미지를 수집하였다. 이후, DBSCAN을 수행하여 엑스트라의 얼굴 이미지 489장을 제외하였으며, 잘못 군집된 이미지 529장을 직접 재분류하였다. 최종적으로 수집한 얼굴 이미지는 총 1,810장이다. 얼굴 이미지를 추가 확보하기 위해 어그멘테이션을 수행하여 최종 54,300장의 얼굴 이미지

Collection @ daejin

를 학습 데이터 세트로 사용하였다.

<표 4-1> 하드웨어 실험 환경

구성요소	사 양	
CPU	Intel(R) Xeon(R) Gold 5120 2.20GHz	
RAM	180GB	
VGA	Tesla V100 SXM2 32GB	Tesla V100 SXM2 32GB
Storage	1,050GB	

<표 4-2> 소프트웨어 실험 환경

구성요소	사 양
OS	Ubuntu 16.04 LTS
CUDA	CUDA 10.1
cuDNN	cuDNN 7.5.1
Lua	LuaJIT
Torch	Torch7
Python	Python 3.6
Tensorflow	tenserflow-gpu 2.1.0

Collection @ daejin

<표 4-3> 영화 '기생충' 등장 인물 정의

배우	캐릭터	중요도	얼굴 이미지 수
Song Kang-Ho	Ki Taek	3	323장
Lee Sun-Kyun	Dong Ik	3	164장
Jo Yeo-Jeong	Yeon Kyo	3	283장
Choi Woo-Sik	Ki Woo	3	353장
Park So-Dam	Ki Jung	3	214장
Jang Hye-jin	Chung Sook	3	204장
Lee Jeon-Geun	Moon Gwang	2	96장
Jeong Ji-So	Da Hye	2	82장
Jung Hyun-Jun	Da Song	2	24장
Park Myeong-Hoon	Geun Se	2	67장

Collection @ daejin

2. 실험 결과

[그림 4-1]은 ‘영역 박스 신뢰도 점수’, ‘영역 면적’, ‘영역과 장면 중심 간 거리’, ‘객체 종류’를 하나씩 고려한 상위 5개 영역 캡션과 핵심 영역 캡션 검출 알고리즘을 적용한 상위 5개 영역 캡션이다. ‘영역 박스 신뢰도 점수’만 고려한 경우 장면에서 등장하는 다양한 객체를 영역 박스가 정확하게 포함하고 있지만 ‘a red bag on the floor’, ‘a red and white bottle’, ‘a window with a frame’ 등 장면에서 상대적으로 중요하지 않은 영역 캡션들도 다수 등장했다. ‘영역 면적’만 고려한 경우 장면에서 차지하는 면적이 큰 객체 위주로 추출되며 장면에서 중요한 ‘white wall in bathroom’, ‘two women holding hands’, ‘man holding a cell phone’과 같은 영역 캡션이 추출되지만 ‘a black and white bag’과 같은 주변 사물에 대한 영역 캡션도 추출되었다. ‘영역과 장면 중심 간 거리’가 가까운 영영 캡션에서는 장면에서 초점을 두고 있는 두 인물에 대해 중점적으로 영역 캡션이 나타났지만, ‘a man wearing a black shirt’, ‘the man has short hair’와 같이 상대적으로 중요하지 않은 세부적인 부분까지 영역 캡션으로 나타났다. ‘객체 종류’만 고려한 경우 중요도가 높은 두 인물의 얼굴을 포함하는 영역 캡션이 우선적으로 추출되고, 이후 사물을 설명하는 캡션이 나타나는 것을 확인할 수 있었다. 이는 ‘the man has short hair’과 같은 얼굴에 대한 세부 영역 캡션을 추출할 수 있으며 중요도가 낮은 사물이나 배경에 대해서는 ‘a red bag on the floor’와 같이 다소 중요하지 않은 영역 캡션이 추출될 수 있다. 핵심 영역 캡션 검출 알고리즘은 4가지 변수를 동시에 고려하며, 추출된 영역 캡션은 ‘two women holding hands’, ‘white wall in bathroom’, ‘man holding a cell phone’, ‘the man has short hair’, ‘woman holding a cell phone’이다. 각 변수를 하나씩 고려한 것보다 핵심 영역 캡

Collection @ daejin

션 검출 알고리즘이 두 인물이 화장실에서 핸드폰을 들고있는 해당 장면을 설명하는데 보다 중요한 영역 캡션을 추출한 것을 확인할 수 있다.

[그림 4-1] 핵심 영역 캡션 검출 알고리즘 결과(상위 5개)

Collection @ daejin

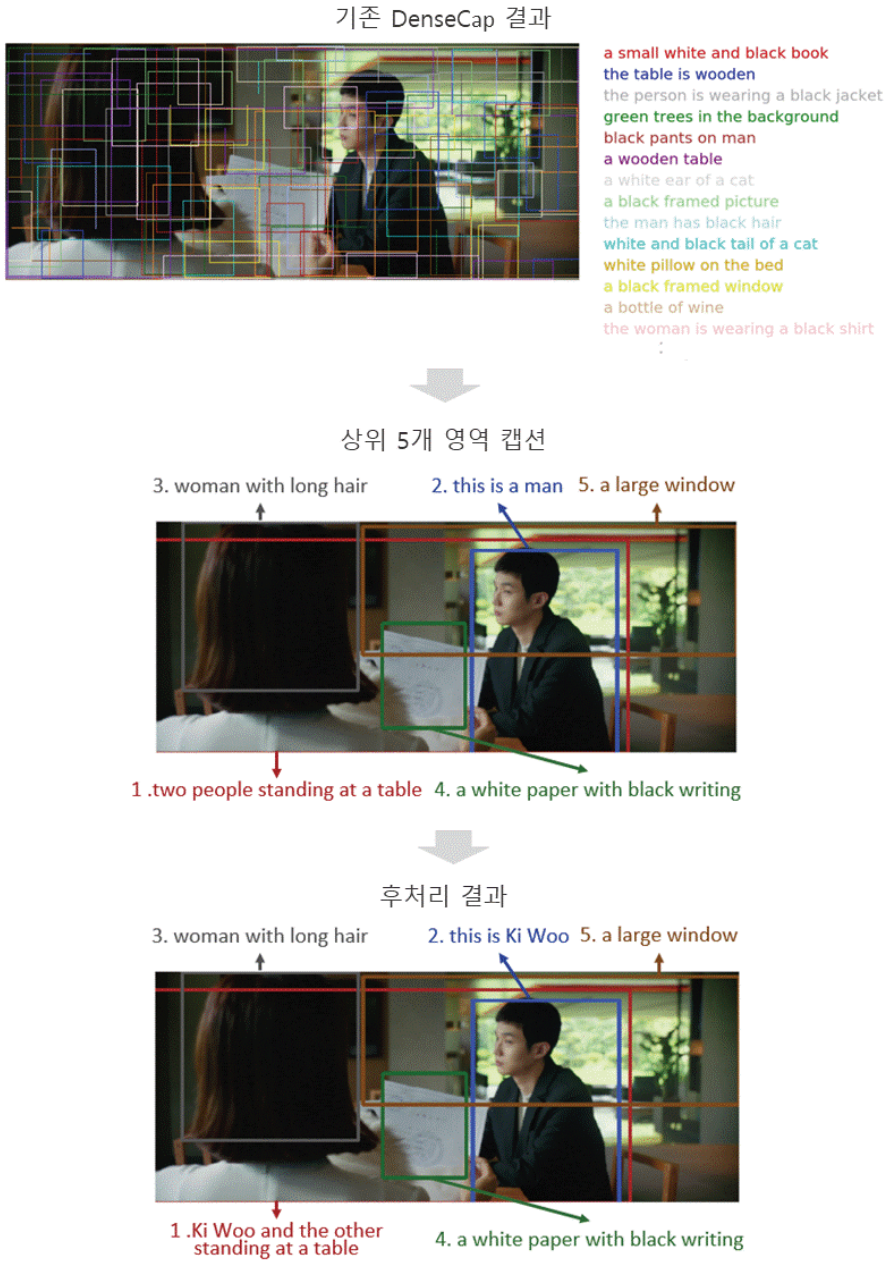
[그림 4-2], [그림 4-3]은 제안 프레임워크를 통해 상위 영역 캡션 5개를 추출하고 후처리를 수행한 결과를 보여준다.

[그림 4-2]에서 추출한 상위 영역 캡션은 ‘two people standing at a table’, ‘this is a man’, ‘woman with long hair’, ‘a white paper with black writing’, ‘a large window’이다. 추출된 캡션을 통해 큰 창문이 있는 집에서 두 인물이 책상에 앉아 검은색 글씨가 쓰인 종이를 들고 있는 것을 알 수 있으며, 해당 장면을 설명하는 의미있는 영역 캡션들이 나타났다고 판단된다. 또한, ‘this is a man’의 영역 캡션이 ‘man’이 사람을 지칭하고 영역 박스가 캐릭터 식별 모듈에서 예측한 캐릭터 ‘Ki Woo’의 얼굴을 70% 이상 포함함으로 ‘man’을 ‘Ki Woo’로 변환하고 관사를 제거하여 ‘this is Ki Woo’로 캡션을 개선한다. ‘two people standing at a table’ 영역 캡션은 두 명의 인물을 나타내지만 포함하는 캐릭터의 얼굴 박스는 한 명이므로 ‘Ki woo and the other standing at a table’로 변환된다.

[그림 4-3]은 후처리 과정을 명확하게 보여준다. 핵심 영역 검출 알고리즘으로 추출된 상위 5개 영역 캡션은 ‘two men standing in a room’, ‘a woman wearing a white shirt’, ‘woman with long hair’, ‘man in black shirt’, ‘man with short brown hair’이다. 영역 캡션 중 네 번째와 다섯 번째 영역 박스의 IoU가 0,2 이상이며 객체가 인물 ‘man’을 나타내므로 영역 박스와 캡션을 통합한다. 마찬가지로, 두 번째와 세 번째 영역 캡션은 객체가 인물 ‘woman’을 나타내므로 영역 박스와 캡션을 통합하며 캐릭터 예측 모듈이 예측한 ‘Yeon Kyo’ 캐릭터 얼굴 박스가 영역 박스에 70% 이상 포함되기 때문에 영역 캡션의 ‘woman’을 ‘Yeon Kyo’로 변환하고 앞에 붙은 관사를 제거하였다. 첫 번째 영역 캡션은 ‘two men standing in a room’으로 두 명의 인물을 나타내고 영역 박스 내 캐릭터 식별 모듈이 예측한 캐릭터가 ‘Yeon Kyo’ 한 명이므로 ‘two men’을 ‘Yeon Kyo and the other’로

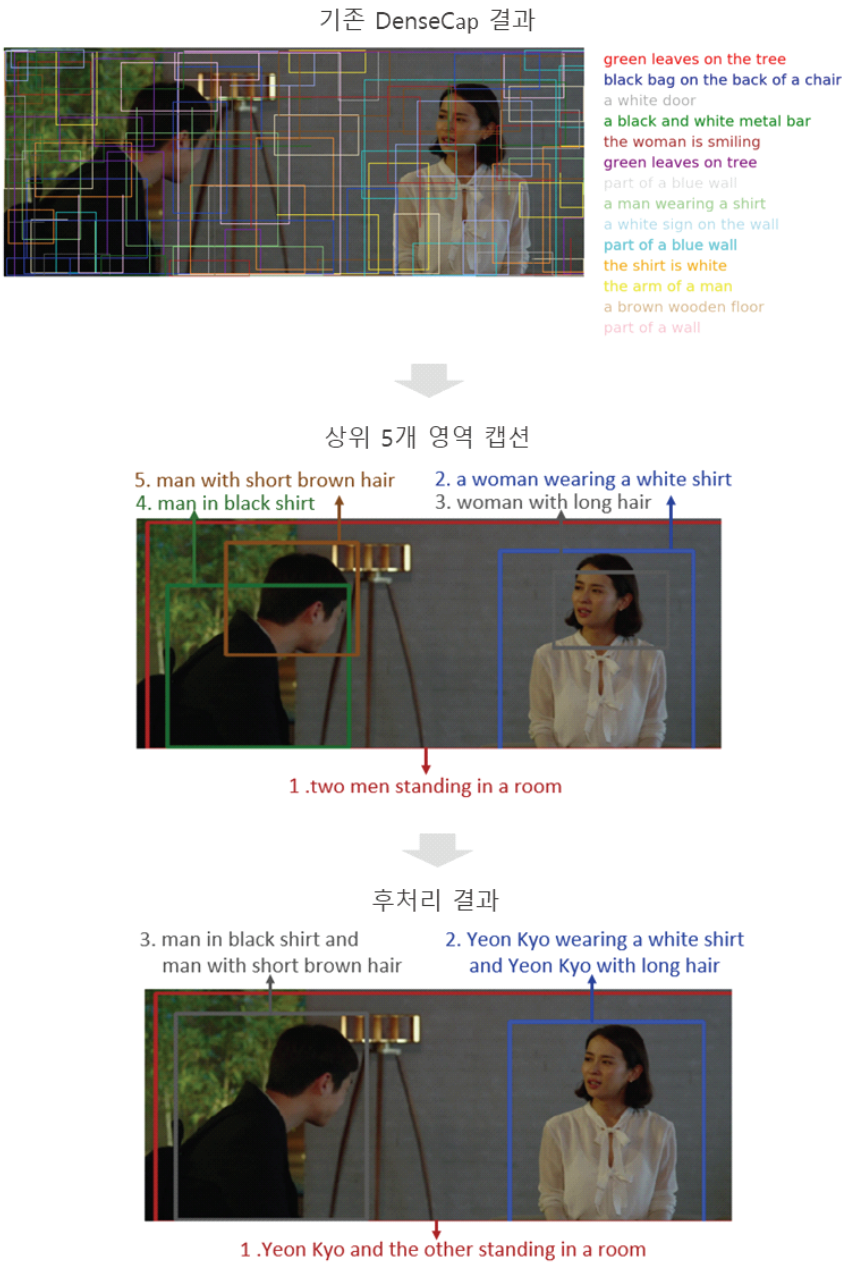
Collection @ daejin

변환한다. 최종 변환된 영역 캡션은 3개로 감소하며 영화 캐릭터 정보를 영역 캡션이 잘 반영하게 되는 것을 확인할 수 있다.



[그림 4-2] 제안 프레임워크 결과

Collection @ daejin



[그림 4-3] 제안 프레임워크 결과(2)

Collection @ daejin

[그림 4-4]는 기존 밀집 캡셔닝 모델 DenseCap과 Dense Relational Captioning이 생성한 영역 캡션과 제안 프레임워크에서 생성한 영역 캡션의 일부를 비교하였다. 해당 이미지에 대해 DenseCap은 95개의 영역 캡션을 생성하였으며, Dense Relational Captioning은 8,930개의 영역 캡션을 생성하였다. 반면, 제안 프레임워크는 동일 객체에 대한 영역 캡션을 통합하여 57개의 영역 캡션을 생성하였다. DenseCap과 Dense Relational Captioning이 생성한 영역 캡션에는 영화 장면을 이해하기에 상대적으로 중요하지 않은 영역 캡션들이 다수 나타난다. 특히, 객체 간 관계에 초점을 맞추는 Dense Relational Captioning은 연관성이 낮은 객체 쌍에 대한 캡션도 생성하기 때문에 DenseCap 보다 높은 비율로 잘못된 영역 캡션을 생성된다. 또한, 기존 밀집 캡셔닝 모델들은 ‘man’, ‘woman’, ‘people’, ‘boy’, ‘girl’ 등 인물을 설명하는 영역 캡션이 어느 캐릭터에 대한 설명인지 특정하기 어렵다. 반면, 제안 프레임워크를 통해 생성된 영역 캡션은 캡션이 나타내는 인물이 식별되면 캐릭터 이름이 캡션에 반영되기 때문에 해당 캡션이 어떤 캐릭터에 대한 설명인지 알 수 있다.

제안 프레임워크는 영역 캡션의 수를 현저히 줄이면서 장면을 이해하는데 상대적으로 중요한 영역 캡션들을 추출한다. 뿐만 아니라, 후처리를 통해 영화를 이해는데 적합한 형태로 변환하여 기존 밀집 캡셔닝 모델보다 개선된 결과를 보였다. 또한, 추출하는 영역 캡션의 수를 제한한다면 더 적은 수의 영역 캡션을 생성할 수 있다.

Collection @ daejin

DenseCap (N=95)	Dense Relational Captioning (N=8,930)	Proposed Framwork (N=57)
a woman sitting on a couch	the talbe has a white	Ki Jung and Ki Taek sitting at a table
the floor is made of wood	the pizza on table	Ki Taek in a blue shirt and Ki Taek holding a glass of wine
a glass of water on the table	the table in room	Ki Jung sitting on a couch and Ki Jung holding a glass and Ki Jung with brown hair
a hand hoding a knife	the woman has a hair	Chung Sook sitting on a couch
green and white bad	the table in front of a white wall	the shirt is green
man in a blue shirt	the table has a black hair	a wooden table and a black and silver table
a woman hoding a glass of wine	the hand has a black hair	a chair in the background and a gray chair
the table is wooden	the glass has a white	a bottle of wine and a black bottle of wine
a chair in the background	the hand on a man	a man holding a glass of wine
a chair in the room	the light on wall	a plate of food and a white plate on a table
...	...	...

[그림 4-4] 기존 밀집 캡셔닝 모델과 제안 프레임워크 캡션 비교

Collection @ daejin

V. 결론

본 연구는 스토리 영상 콘텐츠 중 하나인 영화를 대상으로 영화 자동 이해를 위한 핵심 영역 중심의 이미지 캡셔닝 프레임워크를 제안하였다. 제안 프레임워크에서는 영화의 중요한 요소인 캐릭터 정보를 영역 캡션 중요도 계산과 후처리 과정에 반영하기 위해 캐릭터 식별 모듈을 설계 및 구현하였다. 캐릭터 얼굴 이미지 데이터 세트를 생성하기 위한 알고리즘으로 Haar-Cascade 알고리즘과 HOG 알고리즘을 비교하였으며 HOG 알고리즘의 정확도(98.80%)가 Haar-Cascade 알고리즘 정확도(38.38%) 보다 월등히 높은 것을 보이며 HOG 알고리즘을 채택하여 사용하였다. 또한, 캐릭터의 얼굴 인식을 위한 CNN 모델들을 비교하여 EfficientNet-B5 모델의 얼굴 인식 정확도(99.85%)가 ResNet101(91.61%), VGG-16(85.15) 보다 높은 것을 확인하고 EfficientNet-B5 모델을 채택하여 사용하였다. 제안 프레임워크는 영역 캡션 중요도에 영향을 미치는 변수로 '영역 박스 신뢰도 점수', '영역 면적', '영역과 장면 중심 간 거리', '객체 종류' 4가지를 고려하여 영역 캡션의 중요도를 계산하였다. 기존 밀집 캡셔닝 모델에서 다수 등장하던 불필요한 영역 캡션을 줄이고 영화 장면을 이해하는데 상대적으로 중요한 영역 캡션들이 식별되는 것을 확인하였다. 또한, 동일 객체를 설명하는 영역 캡션을 통합하여 추가적으로 영역 캡션의 수를 줄였으며 영역 캡션이 캐릭터 정보를 반영할 수 있도록 영역 캡션의 인물을 지칭하는 단어를 캐릭터 이름으로 변환하는 과정을 수행하였다. 그 결과, 제안 프레임워크 결과가 기존 밀집 캡셔닝 모듈의 캡션보다 영화 장면을 이해하는데 상대적으로 의미있는 캡션을 추출하였으며, 많은 영화적 요소를 내포하는 것을 확인하였다. 본 연구는 기존 이미지 캡셔닝 도메인에서 다루지 않았던 고난이도 도메인인 스토리 콘텐츠 영상을 대상으로 분석을

Collection @ daejin

진행하였다는 점에서 의의가 있으며, 이러한 연구는 영화의 주석(annotation)과 요약(abstraction), 검색(retrieval), 추천(recommendation) 등 다양한 분야에서 활용될 수 있다. 또한, 실시간성이 보장되어야 하는 응용 분야에서 핵심 영역 캡션만을 고려한다면 연산량과 처리 시간을 줄이는데 기여할 수 있을 것으로 본다.

제안 프레임워크는 개별 영역 캡션과 단일 장면에 대해 분석을 진행하여 객체 및 장면 간의 관계를 고려하지 못한다는 한계를 갖는다. 이러한 한계를 해결하기 위해 장면 간의 관계를 파악하고, 영역 캡션으로부터 객체와 객체 속성, 객체 관계를 정의하여 씬그래프(scene graph)와 온톨로지(ontology)를 구축하는 연구를 진행하고자 한다. 또한, 제안 프레임워크가 다양한 장르의 영화에서 일관된 성능을 보이도록 추가적인 연구를 진행하고자 한다.

Collection @ daejin

참고문헌

[1] 배재윤, “후보군 검출과 HOG 및 SVM을 이용한 선박 검출”, 석사학위 논문, 연세대학교, 2015

[2] Forecast, G. M. D. T. (2019). *Cisco visual networking index: global mobile data traffic forecast update, 2017–2022*.

[3] Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

[4] Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 4565-4574.

[5] Kim, D. J., Choi, J., Oh, T. H., & Kweon, I. S. (2019). Dense relational captioning: Triple-stream networks for relationship-based captioning. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6271-6280.

[6] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... & Bernstein, M. S. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1), 32-73.

[7] Mizuno, K., Terachi, Y., Takagi, K., Izumi, S., Kawaguchi, H., & Yoshimoto, M. (2012). Architectural study of HOG feature extraction processor for real-time object detection. *In 2012 IEEE Workshop on*

Collection @ daejin

Signal Processing Systems, 197-202.

[8] Pérez-Cruz, F., Weston, J., Herrmann, D. J. L., & Scholkopf, B. (2003). Extension of the nu-svm range for classification. *NATO Science Series Sub Series III Computer and Systems Sciences*, 190, 179-196.

[9] Peyre, J., Sivic, J., Laptev, I., & Schmid, C. (2017). Weakly-supervised learning of visual relations. *In Proceedings of the IEEE International Conference on Computer Vision*, 5179-5188.

[10] Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *In Proceedings of the IEEE international conference on computer vision*, 2641-2649.

[11] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[12] Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*.

[13] Tax, D. M., & Duin, R. P. (1999). Support vector domain description. *Pattern recognition letters*, 20(11-13), 1191-1199.

[14] Yang, L., Tang, K., Yang, J., & Li, L. J. (2017). Dense captioning with joint inference and visual context. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2193-2202.

[15] Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic

Collection @ daejin

inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2, 67-78.

Collection @ daejin

Abstract

Dense Captioning Framework Centered on Key Regions for Automatic Understanding of Video Content with Story

Jinsoo So

Department of Computer Engineering

The Graduate School

Daejin University

In the field of deep learning, studies on image classification, object detection, and image captioning have been conducted to visually analyze images and videos. The dense captioning model, one of the image captioning models, extracts various regions in the image and generates region-level captions, but does not provide importance for region captions, so it is difficult to identify region captions that are relatively important for image interpretation. In addition, there is a lack of research on film and drama images, which are story contents.

This study proposes an image captioning framework centered on key

Collection @ daejin

regions for automatic understanding of films, which is one of the story video contents. The proposed framework is based on the DenseCap model, which consists of a character recognition module, a key region caption detection algorithm, and a post-processor. In order to utilize the character information of the film for caption detection and caption improvement in key regions, a character face training data set is created using HOG algorithm and DBSCAN. Based on the learning of the EfficientNet model, a character recognition module that identifies characters in film scenes was designed and implemented. In addition, this study proposes a key region caption detection algorithm that considers four variables that affect the importance of region captions: 'region box confidence score', 'region area', 'distance between region and scene center', and 'object type'. Thereafter, in the post-processor, the region captions for the same object is integrated into one, and the region caption is improved in a form suitable for the film. It was confirmed that the proposed framework effectively identifies important region captions of film scenes and provides more information with fewer region captions than the existing dense captioning model.

Keywords : Image Captioning, Dense Captioning, Face Recognition, Video Content, Film Understanding

Collection @ daejin

Contents

Collection @ daejin

Collection @ daejin

List of Figures

Collection @ daejin

Collection @ daejin

List of Tables

Collection @ daejin