비디오 캡션 생성을 위한 의미 특징 학습과 선택적 주의집중

이수진, 김인철
경기대학교 컴퓨터과학과
e-mail:dltnwls9623@kyonggi.ac.kr, kic@kyonggi.ac.kr

Semantic Feature Learning and Selective Attention for Video Captioning

Sujin Lee, Incheol Kim
Department of Computer Science, Kyonggi University

요 약

일반적으로 비디오로부터 캡션을 생성하는 작업은 입력 비디오로부터 특징을 추출해내는 과정과 추출한 특징을 이용하여 캡션을 생성해내는 과정을 포함한다. 본 논문에서는 효과적인 비디오 캡션 생성을 위한 심층 신경망 모델과 그 학습 방법을 소개한다. 본 논문에서는 입력 비디오를 표현하는 시각 특징 외에, 비디오를 효과적으로 표현하는 동적 의미 특징과 정적 의미 특징을 입력 특징으로 이용한다. 본 논문에서 입력 비디오의 시각 특징들은 C3D, ResNet과 같은 합성곱 신경망을 이용하여 추출하지만, 의미 특징은 본 논문에서 제안하는 의미 특징 추출 네트워크를 활용하여 추출한다. 그리고 이러한 특징들을 기반으로 비디오 캡션을 효과적으로 생성하기 위하여 선택적 주의집중 캡션 생성 네트워크를 제안한다. Youtube 동영상으로부터 수집된 MSVD 데이터 집합을 이용한 다양한 실험을 통해, 본 논문에서 제안한 모델의 성능과 효과를 확인할 수 있었다.

1. 서론

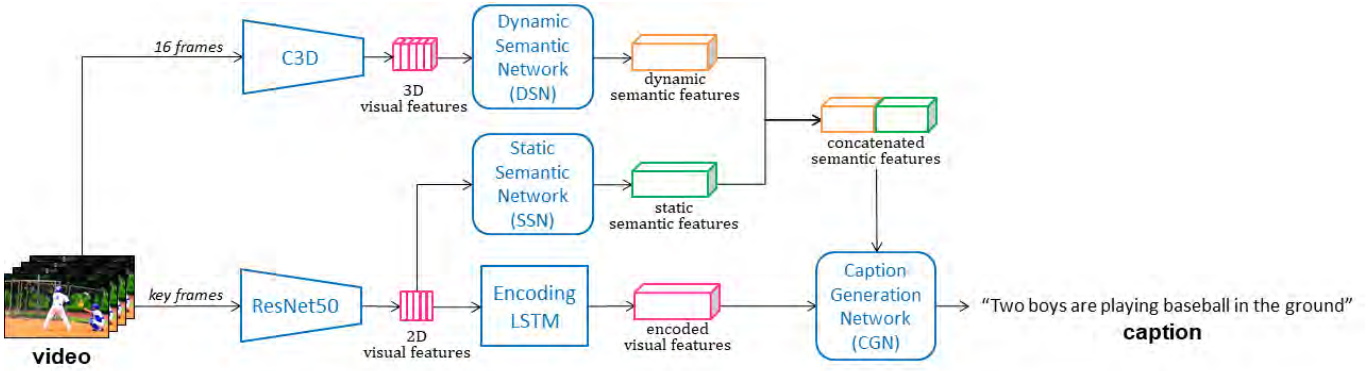
최근 컴퓨터 비전과 자연어 처리, 기계학습 분야에서 인공지능 기술이 발전함에 따라 자연어와 영상을 동시에 처리하는 복합 지능 문제들에 대한 관심이 급증하고 있다. 또한 Youtube, Dailymotion, Netflix 등과 같은 비디오 공유 사이트의 활성화로 인해 비디오 데이터가 증가함에 따라, 영상 뿐 아니라 비디오의 내용을 자동으로 분석하는 문제들에 대한 관심도 증가하고 있다. 대표적인 비디오 기반 복합 지능 문제들로는 비디오 캡션 생성(video captioning), 비디오 기반 질의-응답(video question-answering) 등과 같은 문제들이 있다. 그 중에서도 비디오 캡션 생성은 (그림 1)의 예와 같이 입력 비디오로부터 해당 비디오를 설명하는 자연어 문장을 생성하는 문제를 말한다. 이는 자동 비디오 자막 생성, 비디오 콘텐츠 검색, 비디오 이해 등의 분야에서 활용될 수 있다.

(그림 1) 비디오 캡션 생성의 예

일반적으로 비디오 캡션 생성은 크게 비디오에 대한 특징을 추출해내는 과정(feature extraction)과 이 특징들을 토대로 캡션을 생성하는 과정(caption generation)으로 이루어진다. 최근에는 입력 비디오를 표현하기 위한 특징으로 ResNet, VGG, C3D와 같은 합성곱 신경망(Convolutional Neural Network, CNN)을 이용하여 추출한 시각 특징(visual feature) 뿐 아니라 의미 특징(semantic feature)을 활용하는 연구들[1, 2, 3]이 활발하게 진행되고 있다. 의미 특징이란 비디오 내의 행위, 물체와 같은 속성을 나타내는 단어를 말한다. 이를 활용하기 위해서는 의미 특징을 추출하는 모델과 해당 모델을 학습시킬 데이터 집합이 필요하다. 한편 캡션을 생성하기 위한 모델로는 순환 신경망(Recurrent Neural Network, RNN)을 이용하는데, 순환 신경망의 종류 중 하나인 LSTM(Long Short-Term Memory)이 주로 사용된다. 의미 특징을 캡션 생성에 이용하는 기존의 방법으로는 단순히 매 시간 단계(timestep)마다 캡션을 생성하는 순환 신경망의 입력으로 사용하는 방법, 순환 신경망의 내부 파라미터 가중치(weight)로 사용하는 방법[1], 의미 특징에 주의집중(attention)을 적용하여 사용하는 방법[2], 추출된 의미 특징을 임베딩(embedding)하여 사용하는 방법[3] 등이 있었다. 또한 캡션을 생성하는 모델로 하나의 층으로 이루어진 LSTM 모델을 사용하는 것이 아니라 계층적 구조의 LSTM 모델을 사용하는 연구[4]도 진행되고 있었다. 하지만 기존 연구들에서는 캡션 데이터 집합의 동사, 명사로부터 수집된 의미 특징들을 구별하지 않았고, 의미 특징을 이용하여 캡션을 생성할 때는 비교적 단순한 LSTM 모델을 사용하였다는 한계점[1, 2, 3]이 있었다.

본 논문에서는 비디오 캡션 생성에 효과적인 심층 신경망 모델과 학습 방법을 제시한다. 본 논문에서 제안하는 비디오 캡션 생성 모델에서는 입력 비디오를 효과적으로 표현하기 위하여 시각 특징뿐만 아니라, 비디오를 표현하는 의미 특징을 함께 이용한다. 본 논문에서 시각 특징들은 ResNet 합성곱 신경망을 이용하여 추출하지만, 의미

* 본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음 (IITP-2017-0-01642)



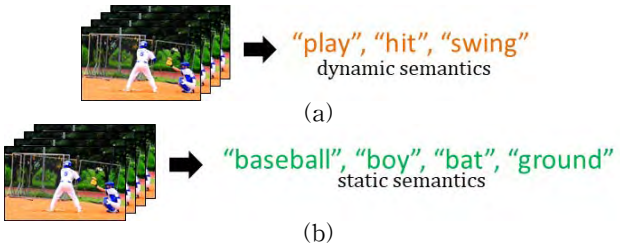
(그림 2) 비디오 캡션 생성 모델

특징은 본 논문에서 제안하는 의미 특징 네트워크를 이용하여 추출한다. 특히 의미 특징을 비디오 내의 행위를 나타내는 동적 의미 특징(dynamic semantic feature)과 비디오 내의 물체, 사람, 배경 등을 나타내는 정적 의미 특징(static semantic feature)으로 구별하여 추출한 뒤 입력 특징으로 함께 이용한다. 그리고 이러한 특징들을 기반으로 효과적으로 캡션을 생성하기 위하여 본 논문에서 제안하는 선택적 주의집중 캡션 생성 네트워크를 이용한다. 선택적 주의집중 캡션 생성 네트워크는 매 시간단계마다 입력된 의미 특징들 중 어떤 의미 특징에 집중할 것인지를 판단하여 캡션을 생성한다. 본 논문에서 제안하는 모델의 성능과 효과를 분석하기 위해 Youtube 동영상으로부터 수집된 MSVD(Microsoft Video Description) 데이터 집합을 이용한 다양한 실험을 수행하고 그 결과를 소개한다.

2. 비디오 캡션 생성 모델

2.1 모델 개요

본 논문에서는 효과적인 비디오 캡션 생성을 위하여 비디오를 표현하는 시각 특징 외에, 의미 특징을 이용한 캡션 생성을 제안한다. 합성곱 신경망을 이용하여 추출된 시각 특징은 입력 비디오 내의 사람, 물체, 배경, 행위 등의 속성들을 일괄적으로 함축하여 표현하는 특징이다. 이러한 특징에 내재된 상위 단계(high-level)의 의미 있는 특징들을 보다 직접적인 형태로 캡션 생성에 이용하기 위하여 단어의 형태로 추출하고, 이를 의미 특징으로 활용한다. 따라서 캡션 생성 모델은 입력 비디오를 나타내는 단어로 표현된 의미 특징들을 통해 입력 비디오를 보다 효과적으로 이해할 수 있게 된다. 본 논문에서는 특히 의미 특징을 동적 의미 특징과 정적 의미 특징으로 나누어 추출한 뒤 이를 선택적 주의집중 캡션 생성 네트워크의 추가적인 특징으로 이용한다.

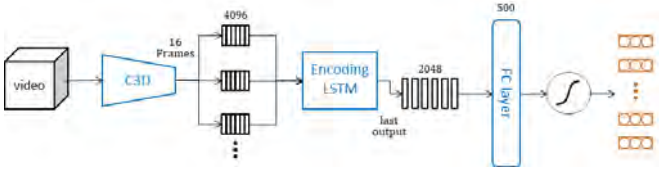


(그림 3) 동적 의미 특징과 정적 의미 특징의 예

동적 의미 특징이란 (그림 3)의 (a)와 같이 비디오 내의 행동에 해당하는 특징을 의미한다. 이와 다르게 정적 의미 특징이란 (그림 3)의 (b)와 같이 비디오 내의 물체, 사람, 배경 등에 해당하는 특징을 의미한다. 즉, 캡션 문장에서 동사에 해당하는 단어는 동적인 의미 특징이고 명사에 해당하는 단어는 정적인 의미 특징이라 할 수 있다. 본 논문에서 제안하는 모델의 전체 구조는 (그림 2)와 같다. 먼저 캡션을 생성하기 위해 필요한 시각 특징들을 미리 학습된 ResNet과 C3D를 이용하여 추출한다. 추출된 시각 특징들은 2.2절에서 소개되는 동적 의미 특징 네트워크(Dynamic Semantic Network, DSN)와 정적 의미 특징 네트워크(Static Semantic Network, SSN)의 입력으로 주어진다. 이후 각 의미 특징 네트워크로부터 동적 의미 특징과 정적 의미 특징을 추출한다. 추출된 의미 특징들은 단순 결합(concatenate)되어 매 시간단계마다 2.3절에서 소개되는 선택적 주의집중 캡션 생성 네트워크(Caption Generation Network, CGN)의 입력으로 주어진다. ResNet을 통해 추출된 시각 특징은 정적 의미 특징 네트워크의 입력 외에도 시각 특징을 인코딩하는 LSTM의 입력으로 주어지고, 인코딩 LSTM의 마지막 출력은 캡션 생성 네트워크의 초기화에 주어진다. 캡션 생성 네트워크는 매 시간단계마다 어떠한 의미 특징에 집중할 것인지를 판단하여 단어들의 확률 분포를 계산한다. 이후 출력된 단어들의 확률 분포를 통해 캡션을 생성한다.

2.2 의미 특징 학습

의미 특징을 활용한 캡션 생성을 위해서는 입력 동영상으로부터 의미 특징을 알아내야 한다. 앞서 소개한바와 같이 의미 특징은 행위를 나타내는 동적인 특징과 물체, 사람, 배경 등을 나타내는 정적인 특징으로 나눌 수 있다. 이러한 동적 특징과 정적 특징에는 분명한 차이점이 존재한다. 동적 특징은 비디오의 한 장면, 즉 한 프레임으로는 알아내기 어렵고, 비디오를 일정 시간동안 관찰함으로써 알아낼 수 있다. 이와 달리 정적 특징은 비디오의 한 장면에 등장하는 물체, 사람, 배경 등에 해당하므로 비디오의 한 프레임을 관찰함으로써 알아낼 수 있다. 따라서 본 논문에서는 이를 의미 특징 네트워크에 적용하기 위하여 동적 의미 특징과 정적 의미 특징을 구별하여 추출하고, 이를 다중 범주 분류(multi label classification) 문제로 간주한다. 특히 동적 의미 특징의 경우 비디오의 시간적, 공간적 특징을 모두 효과적으로 표현한 시각 특징을 활용하여 추출하고, 정적 의미 특징의 경우 비디오의 공간적 특징을 효과적으로 표현한 시각 특징을 활용하여 추출한다.



(그림 4) 동적 의미 특징 네트워크

본 논문에서 제안하는 동적 의미 특징 네트워크는 (그림 4)와 같다. 먼저 비디오의 시간적, 공간적 특징을 효과적으로 표현하는 시각 특징을 활용하기 위해 미리 학습된

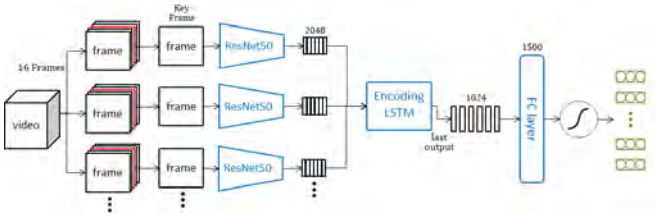
C3D 합성곱 신경망으로부터 (식 1)과 같이 비디오의 16프레임마다 시각 특징을 추출한다. (식 1)의 v_i는 비디오의 한 프레임을 나타내고, n_v는 비디오의 총 프레임 수를, $\frac{n_v}{16}$은 비디오를 16프레임씩 나눴을 때의 나눠진 클립(clip)의 총 개수를 나타낸다. 이후 추출된 시각 특징은 LSTM 순환 신경망 모델을 이용하여 (식 2)와 같이 인코딩된다. c_t는 현재 시간단계(t)에서 인코딩될 한 클립에 해당하는 시각 특징을 나타내고 h_{t-1}은 LSTM의 이전 은닉 상태를 나타낸다.

$$c_i = \mathrm{C3D}(v_{i:i+16}),\ i \in \left\{0,1,...,\frac{n_v}{16}\right\} \quad \text{(식 1)}$$

$$e = \mathrm{LSTM}(c_t, h_{t-1}) \quad \text{(식 2)}$$

이후 인코딩 된 시각 특징(e)으로부터 완전 연결 계층과 활성 함수인 시그모이드 함수(sigmoid)를 통해 (식 3)과 같이 동적 의미 특징의 확률 분포(p_d)를 알아내는데, W_d는 학습해야할 가중치를, b_d는 바이어스를 나타낸다.

$$p_d = sigmoid(W_d \cdot e + b_d) \quad \text{(식 3)}$$



(그림 5) 정적 의미 특징 네트워크

본 논문에서 제안하는 정적 의미 특징 네트워크는 (그림 5)와 같다. 먼저 비디오의 공간적인 특징을 효과적으로 표현하는 시각 특징을 활용하기 위해 미리 학습된 ResNet 합성곱 신경망으로부터 시각 특징을 추출한다. (식 4)와 같이 비디오를 16프레임씩 나눈 뒤 가운데 프레임에 해당하는 8번째 프레임들로부터 시각 특징(r_i)을 추출하고, 이를 LSTM을 이용하여 (식 5)와 같이 인코딩한다. 이후 완전 연결 계층과 시그모이드 함수를 이용하여 정적 의미 특징의 확률 분포(p_s)를 알아낸다.

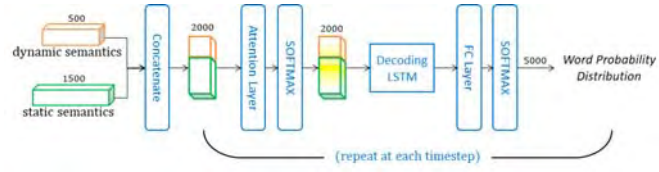
$$r_i = \mathrm{ResNet}(v_{i\times 8+16}),\ i \in \left\{0,1,...,\frac{n_v}{16}\right\} \quad \text{(식 4)}$$

$$e = \mathrm{LSTM}(r_t, h_{t-1}) \quad \text{(식 5)}$$

$$p_s = sigmoid(W_s \cdot e + b_s) \quad \text{(식 6)}$$

2.3 선택적 주의집중 캡션 생성

본 논문에서는 의미 특징을 활용한 효과적인 캡션 생성을 위해 (그림 6)과 같은 선택적 주의집중 캡션 생성 네트워크를 제안한다.



(그림 6) 캡션 생성 네트워크

캡션 생성 네트워크는 매 시간단계마다 동적 의미 특징과 정적 의미 특징을 입력으로 받아 단어들의 확률 분포를 알아낸다. 동적 의미 특징과 정적 의미 특징은 단순 결합되어 주의집중 계층(attention layer)의 입력으로 이용된다. 일반적으로 캡션을 생성할 때 현재 생성될 단어가 명사(noun)라면 비디오 내의 물체(object)에, 동사(verb)라면 비디오 내의 행위(behavior)에 집중하여 생성하는 것이 보다 효과적인 방법일 것이다. 본 논문에서는 이를 캡션 생성 네트워크에 적용하기 위하여 캡션을 생성할 때 현재 시간단계에서 어떠한 의미 특징에 보다 집중해야 하는지를 주의집중 계층을 통해 판단한다. 주의집중 계층에서는 현재 시간단계(t)에서 어떤 특정 의미 특징에 집중할 것인지를 나타내는 가중치(W_a)가 적용된 의미 특징을 계산한다. 이러한 가중치가 적용된 의미 특징(a_t)은 (식 7)을 통해 계산되는데, s_t는 입력으로 주어진 의미 특징을, b_a는 바이어스(bias)를 나타낸다.

$$a_t = softmax(W_a \cdot s_t + b_a) \quad \text{(식 7)}$$

이후 변환된 의미 특징들은 디코딩(decoding) LSTM의 입력으로 주어진다. 디코딩 LSTM은 입력된 의미 특징들로부터 문장 구조를 학습하여 (식 8)과 같이 현재 시간단계에서 어떠한 단어를 생성해야 하는지를 나타내는 상태 값을 출력한다. 디코딩 LSTM의 초기 은닉 상태($h_{t=0}$)는 시각 특징을 인코딩하는 인코딩 LSTM의 마지막 은닉 상태 값으로 초기화된다.

$$h_t = \mathrm{LSTM}(a_t, h_{t-1}) \quad \text{(식 8)}$$

디코딩 LSTM의 출력은 다시 완전 연결 계층(fully connected layer)의 입력으로 주어진다. 완전 연결 계층에서는 (식 9)와 같이 현재 시간단계(t)에서 어떤 단어가 적절할지를 나타내는 확률 분포(p_t)를 계산하는데, W_p는 학습해야할 가중치를, h_t는 디코딩 LSTM으로부터 주어진 입력을, b_p는 바이어스를 나타낸다.

$$p_t = softmax(W_p \cdot h_t + b_p) \quad \text{(식 9)}$$

매 시간단계마다 입력 의미 특징들에 대한 주의집중을 계산하고, 디코딩 LSTM을 거쳐 완전 연결 계층을 통해 단어의 확률 분포가 출력되는 작업이 반복된다. 이후 첫 번째 단어부터 문장의 끝을 나타내는 단어인 '<EOS>' 전까지를 캡션으로 생성하게 된다.

3. 구현 및 실험

3.1 데이터 집합

본 논문에서 제안하는 캡션 생성 모델을 학습하고 평가하기 위하여 Youtube 비디오로부터 수집된 비디오 캡션 데이터 집합인 MSVD 데이터 집합을 사용하였다. MSVD 데이터 집합은 1970개의 Youtube 비디오 클립과 이에 해당하는 약 80,000개의 캡션 문장으로 이루어져있으며 학습, 검증, 테스트 집합이 각각 1200, 100, 670개의 동영상으로 나뉘어져 있다.

본 논문에서 제안하는 의미 특징 네트워크를 학습하기 위해서는 학습에 사용할 데이터 집합이 필요하다. 본 논문에서는 의미 특징 학습을 위한 데이터 집합을 수집하기 위하여 MSVD 비디오 캡션 데이터 집합을 사용하였다. 먼저, MSVD 데이터 집합의 캡션 문장들을 NLTK (Natural Language Toolkit)의 POS(Part-Of-Speech) 태그 기능을 이용하여 명사, 동사로 분리하고 명사의 복수형이나 동사의 과거형, 진행형 등의 시제를 NLTK의 레머타이즈(lemmatize) 기능을 활용하여 기본형으로 변환하였다. 이후 분리된 동사들 중 등장 빈도수가 높은 500개를 선택하여 동적 의미 특징의 라벨 데이터(label data)로 구성하였고, 분리된 명사들 중 등장 빈도수가 높은 1500개를 선택하여 정적 의미 특징의 라벨 데이터로 구성하였다. 동적 의미 특징의 라벨 데이터 중 특정 동사가 비디오 내의 캡션에 포함된 동사이면 이 동사에 대한 해당 비디오의 라

벨을 1로 표기(labeling)하고, 포함되지 않은 동사이면 0으로 표기하여 동적 데이터 집합을 구성하였다. 이와 동일한 방식으로 정적 데이터 집합도 구성하였다. 각 비디오는 구성한 데이터 집합 내의 약 7개의 명사와 약 3개의 동사를 가진다. 의미 특징 데이터 집합도 MSVD 캡션 데이터 집합과 동일하게 학습, 검증, 테스트 집합을 각각 1200, 100, 670개로 구성하였다.

3.2 모델 학습

실험을 위해 Ubuntu 14.04 LTS 환경에서 Python 딥러닝 라이브러리인 Keras를 이용하여 본 논문에서 제안하는 모델을 구현하였다. 입력으로 사용되는 비디오들의 길이는 16프레임으로 이루어진 40개의 클립으로 구성되도록 일정하게 샘플링(uniform sampling)되었다. 의미 특징 네트워크의 경우 모델 최적화 알고리즘은 Adam을, 손실 함수는 (식 10)에 표현된 이진 크로스엔트로피(binary cross-entropy)를 사용하여 학습하였다. y는 실제 정답, $\tilde{y}$는 예측치를 나타낸다.

$$L_{binary} = -[y\log\tilde{y} + (1-y)\log(1-\tilde{y})] \quad \text{(식 10)}$$

의미 특징 추출 네트워크의 학습이 완료되면 캡션 데이터 집합의 모든 비디오로부터 의미 특징을 미리 추출해놓은 후, 캡션 생성 네트워크의 입력으로 사용한다. 캡션 생성 네트워크의 경우 모델 최적화 알고리즘은 RMSprop을, 손실 함수는 (식 11)에 표현된 범주별 크로스엔트로피(categorical cross-entropy)를 사용하였다.

$$L_{categorical} = -\frac{1}{n}\sum_{x}[y\log\tilde{y} + (1-y)\log(1-\tilde{y})] \quad \text{(식 11)}$$

의미 특징 추출 모델의 경우 일괄 처리량(batch size)은 32, 반복 횟수(epoch)는 500으로 설정하였고, 캡션 생성 모델의 경우 일괄 처리량은 25, 반복 횟수는 50으로 설정하여 학습을 수행하였다.

3.3 실험과 평가

첫 번째 실험에서는 각 의미 특징이 캡션 생성 성능에 미치는 영향을 파악하기 위한 실험을 진행하였다. 이를 위해 캡션 생성 네트워크는 본 논문에서 제안한 선택적 주의집중 캡션 생성 네트워크로 고정하여 사용하고, 입력 특징에 차이를 두어 실험을 진행하였다. 캡션 생성 네트워크의 성능을 평가하기 위한 척도로는 일반적으로 캡션 생성 평가 지표(metric)로 사용되는 BLEU@N, CIDEr-D를 사용하였다. 모든 평가 지표는 Microsoft COCO evaluation server에서 제공되는 코드를 사용하여 계산되었다. <표 1>의 CGN은 의미 특징을 사용하지 않고 시각 특징만을 이용하여 캡션을 생성한 경우, DSN+CGN은 동적 의미 특징만을 이용한 경우, SSN+CGN은 정적 의미 특징만을 이용한 경우, DSN+SSN+CGN은 동적, 정적 의미 특징을 모두 사용한 경우를 나타낸다.

<표 2> 의미 특징 네트워크 간 성능 비교

Models	B@1	B@2	B@3	B@4	CIDEr
CGN	66.1	47.8	37.1	26.5	26.4
DSN+CGN	76.0	58.1	45.7	35.8	50.0
SSN+CGN	78.8	63.4	51.4	41.4	77.8
DSN+SSN+CGN	**84.8**	**70.8**	**60.0**	**50.0**	**94.3**

<표 1>의 결과를 살펴보면 의미 특징을 사용한 모델의 성능이 더 뛰어난 것을 확인할 수 있었다. 특히 동적 의미 특징만을 활용했을 때보다 정적 의미 특징만을 활용한 모델이 더 성능이 좋았는데, 이는 동적 의미 특징이 비디오 내의 행위만을 나타낸 특징인 반면 정적 의미 특징은 비디오 내의 사람, 물체, 배경 등을 나타내기 때문에 보다 비디오를 잘 표현하는 특징이기 때문이다. 또한 두 가지의 의미 특징을 모두 활용한 모델의 성능이 가장 좋았는데, 이를 통해 두 가지의 의미 특징이 독립적으로 캡션 생성 성능에 기여했다는 것을 확인할 수 있다.

두 번째 실험에서는 본 논문에서 제안한 캡션 생성 모델인 SeFLA의 성능을 비교 평가하기 위한 실험을 진행하였다. <표 2>는 본 논문의 캡션 생성 모델인 SeFLA의 성능과 기존 연구들의 성능을 비교한 결과이다. <표 2>의 SCN은 Gan의 연구에서 제안된 모델이고 LSTM-TSA는 Pan의 연구, hLSTMat는 Song의 연구에서 제안된 모델이다. SCN과 LSTM-TSA는 의미 특징을 사용한 모델이고, hLSTMat는 의미 특징을 사용하지 않았지만 캡션을 생성하기 위한 순환 신경망으로 주의집중 기반 계층적 LSTM을 사용한 모델이다.

<표 3> 기존 모델들과의 성능 비교

Models	B@1	B@2	B@3	B@4	CIDEr
SCN [1]	-	-	-	51.1	77.7
LSTM-TSA [2]	82.8	72.0	62.8	52.8	74.0
hLSTMat [4]	82.9	**72.2**	**63.0**	**53.0**	73.8
SeFLA	**84.8**	70.8	60.0	50.0	**94.3**

<표 2>의 결과를 살펴보면 본 논문에서 제안한 SeFLA가 BLEU@1, CIDEr에서 84.8%, 85.7%로 기존 연구들보다 각각 1.9%, 10.3% 더 우수함을 알 수 있었다. BLEU@2, 3, 4의 성능은 기존 연구들보다 미흡했는데, 본 논문에서 제안한 SeFLA가 단어 1개씩은 더 잘 맞추지만 단어 여러 개를 연속으로 맞추는 것은 상대적으로 부족함을 알 수 있다. 이는 SeFLA가 의미 특징의 도움으로 캡션 내의 명사나, 동사는 잘 생성하지만 상대적으로 문장 구조상 필요한 조사, 전치사 등을 잘 생성해내지 못하였음을 알 수 있다. 이는 학습 데이터의 부족으로 캡션 생성 네트워크의 LSTM의 문장 구조에 대한 학습이 충분히 이루어지지 않았기 때문이다. 전반적으로 봤을 때, 본 논문에서 제안한 SeFLA가 의미 특징을 효과적으로 적용하여 캡션을 생성해내고 있음을 알 수 있다.

4. 결론

본 논문에서는 비디오 캡션 생성에 효과적인 심층 신경망 모델을 제시하였다. 본 논문에서 제안하는 캡션 생성 모델은 입력 비디오로부터 합성곱 신경망을 이용하여 추출한 시각 특징 이외에도 의미 특징 추출 네트워크로부터 추출한 의미 특징을 입력 특징으로 이용한다. 또한 의미 특징을 효과적으로 캡션 생성에 적용하기 위하여 의미 특징에 주의집중을 적용한 선택적 주의집중 캡션 생성 네트워크를 제안하였다.

참고문헌

[1] Z. Gan, C. Gan, and X. He, et. al., “Semantic Compositional Networks for Visual Captioning,” Proc. of CVPR-17, 2017.

[2] Y. Pan, T. Yao, and H. Li, et. al., “Video Captioning with Transferred Semantic Attributes,” Proc. of CVPR-17, 2017.

[3] Y. Yu, H. Ko, and J. Choi, et. al., “End-To-End Concept Word Detection for Video Captioning, Retrieval, and Question Answering,” Proc. of CVPR-17, 2017.

[4] J. Song, Z. Guo, and L. Gao, et. al., “Hierarchical LSTM with Adjusted Temporal Attention for Video Captioning,” Proc. of IJCAI-17, 2017.