Journal of KIIT. Vol. 19, No. 5, pp. 19-29, May 31, 2021. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2021.19.5.19

폭우 환경에서의 의미적 시각 특징 정합 기반의 이미지 캡셔닝

여풍휘*, 손창환**

Image Captioning via Semantic Visual Feature Matching in Heavy Rain Condition

Pung-Hwi Ye* and Chang-Hwan Son**

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임
(No. 2020R1A2C1010405)

요 약

입력 영상으로부터 문장을 생성하는 이미지 캡셔닝 기법은 청명한 날씨에서 촬영된 고화질의 영상을 대상으로 개발되고 있다. 하지만 폭우, 폭설, 안개와 같은 악천후 환경에서는 가시성 저하, 빗줄기 및 눈송이 결함으로 촬영된 영상의 화질은 열화된다. 이는 객체의 유용한 특징 추출을 방해하며 이미지 캡셔닝의 성능 저하를 유발한다. 이런 문제를 극복하고자 본 연구에서는 폭우 환경에서도 정확한 문장을 생성할 수 있는 이미징 캡셔닝 기법을 제안하고자 한다. 먼저, 물리적 폭우 모델링 기반의 폭우 영상 화질 개선 네트워크를 제시한 후, 물리적 폭우 모델링의 역과정을 통해 폭우 영상의 화질을 복구하고자 한다. 그리고 복구 영상과 원본 영상 간의 시각 특징 벡터를 정합할 수 있는 인코더를 제안하고 복구 영상의 의미적 시각 표현력을 제고하고자 한다. 실험 결과를 통해, 제안한 이미지 캡셔닝 기법이 기존 이미지 캡셔닝 기법에 비해 폭우 환경에서의 문장 생성의 정확도를 높일 수 있음을 보이고자 한다.

Abstract

Image captioning methods for generating sentences from input images are being developed for high-quality images captured in clear weather. However, in bad weather conditions such as heavy rain, heavy snow, and fog, the quality of the captured images deteriorates due to poor visibility, rain and snow artifacts. This hinders the extraction of useful features of the object and causes the performance of image captioning to deteriorate. To overcome this problem, this study proposes an image captioning method that can generate accurate sentences even in heavy rain conditions. First, an image quality improvement network based on physical heavy rain modeling is presented, and then the image quality of heavy rain images is restored through the reverse process of physical heavy rain modeling. In addition, an encoder capable of matching visual feature vectors between the restored image and the original image is proposed, and the semantic visual feature expression power of the restored image is improved. Through experiments, it is confirmed that the proposed image captioning method can increase the accuracy of sentence generation in a heavy rain condition, compared to the existing image captioning method.

Keywords
image captioning, rain removal, deep learning

* 군산대학교 소프트웨어융합공학과
- ORCID: https://orcid.org/0000-0002-2446-7022
** 군산대학교 소프트웨어융합공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0001-7077-3074

· Received: Mar. 26, 2021, Revised: Apr. 20, 2021, Accepted: Apr. 23, 2021
· Corresponding Author: Chang-Hwan Son
Department of Software Convergence Engineering, Republic of Korea,
Tel.: 82+63-469-8915, Email: cson@kunsan.ac.kr

Ⅰ. 서 론

딥 러닝의 등장과 함께 객체 검출 및 영상 분류와 같은 컴퓨터 비전 분야는 급속도로 발전되고 있다. 그리고 딥 러닝이 영상이나 텍스트와 같은 입력 데이터와는 무관하게 유용한 특징을 자동으로 추출할 수 있어서 학문 간의 경계가 허물어지고 멀티모달 딥 러닝(Multimodal deep learning)에 대한 관심이 증폭되고 있다. 멀티모달 딥 러닝의 대표적인 분야로는 이미지 캡셔닝(Image captioning)을 들 수 있다. 이미지 캡셔닝이란 입력 영상의 배경이나 객체의 행동을 글로써 묘사하는 기술을 말한다. 쉽게 말하면, 입력 영상으로부터 문장을 생성하는 기술이다. 이러한 이미지 캡셔닝 기술은 기존의 객체 존재 유무만을 판별하는 객체 검출보다 객체와 객체 또는 객체와 배경과의 상호관계를 묘사할 수 있는 한 단계 더 고도화된 인공지능 분야라 볼 수 있다. 이러한 이미지 캡셔닝 기술은 영상 자막 생성, 비디오 줄거리 요약, 자동차 주행 중의 상황 설명, 시각 장애인 스마트 안경, 문장 기반 영상 검색, 감시 카메라 상황 인식 분야 등에 적용이 가능하다.

하지만 기존의 이미지 캡셔닝 기술은 청명한 날씨에 촬영된 고해상도의 실외 영상이나 실내 환경에서 촬영된 선명한 영상을 대상으로 개발되고 있다[1]-[3]. 즉, 다양한 기상 환경을 반영한 야외 전용 이미지 캡셔닝 기술에 대한 개발은 거의 전무한 상태이다. 폭우, 폭설, 안개, 저조도와 같은 악천후 환경에서 촬영된 영상은 잡음, 안개, 빗줄기, 눈송이와 같은 결함(Artifacts)을 포함한다. 이러한 결함은 영상 화질 저하를 수반하기 때문에 결국 이미지 캡셔닝의 성능 저하도 뒤따를 수밖에 없다[4][5]. 특히 폭우 환경 경우에는 빗줄기(Rain streak)와 빗줄기 축적(Rain accmulation) 현상 때문에 현격한 가시성 저하가 발생한다. 이로 인해 객체 특징 추출의 어려움이 발생하게 되고 결국 생성된 문장의 정확도를 떨어뜨리게 된다.

본 연구에서는 이런 현안을 극복하기 위해 폭우 환경에서 이미지 캡셔닝의 성능을 개선할 수 있는 딥 러닝 모델을 소개하고자 한다. 첫째, 폭우 환경에서의 이미지 캡셔닝의 성능은 결국 촬영된 입력 영상에서 빗줄기 패턴과 빗줄기 축적 현상을 제거해서 고품질의 영상을 복원하는 과정을 요구한다. 이를 위해, 본 논문에서는 물리적 빗줄기 모델링에 기반한 초기 복원 서브네트워크를 제시하고자 한다. 둘째, 이미지 캡셔닝의 인코더 부분, 즉 입력 영상에서 의미적 시각 정보를 추출해서 특징 벡터로 변환하는 과정을 개선하고자 한다. 비록 초기 복원 서브네트워크를 적용할지라도 초기 복원 영상의 화질은 선명한 날씨에서 촬영한 수준보다는 떨어진다. 따라서 초기 복원 영상에서 추출된 특징 벡터의 의미적 표현력(Semantic representation)을 제고하기 위한 의미적 시각 특징 정합 기반의 인코더를 제안하고자 한다. 셋째, 성능 평가를 통해 제안한 기법이 폭우 환경에서 촬영된 열화 영상을 대상으로 기존의 이미지 캡셔닝 기법보다 더 정확한 문장을 생성할 수 있음을 보이고자 한다.

이 논문의 구성은 다음과 같다. 먼저 제 2장에서 기존의 어텐션 기반의 이미지 캡셔닝 모델을 소개한 후, 제 3장에서는 제안한 폭우 환경에서의 의미적 시각 특징 정합 기반의 이미지 캡셔닝에 대해 자세히 설명하고자 한다. 그리고 제 4장에서는 제안한 기법의 성능을 평가하고 마지막 제 5장에서 결론을 맺고자 한다.

Ⅱ. 기존 이미지 캡셔닝 모델

최근 이미지 캡셔닝 기법은 합성곱 신경망(Convolutional neural networks)과 순환 신경망(Recurrent neural networks)으로 구성된 어텐션 딥 러닝 모델에 기반을 둔다[6]. 합성곱 신경망은 인코더 부분으로써, 입력 영상에서 의미적 시각 특징을 추출하는 역할을 담당하고 순환 신경망은 시각 특징 추출과 입력 단어와의 연관성을 모델링하여 최종 문장을 생성하는 디코더 역할을 한다.

2.1 인코더

인코더는 그림 1과 같이 합성곱 신경망으로 구성되며 입력 영상에서 의미적 시각 특징을 추출하는 역할을 한다.

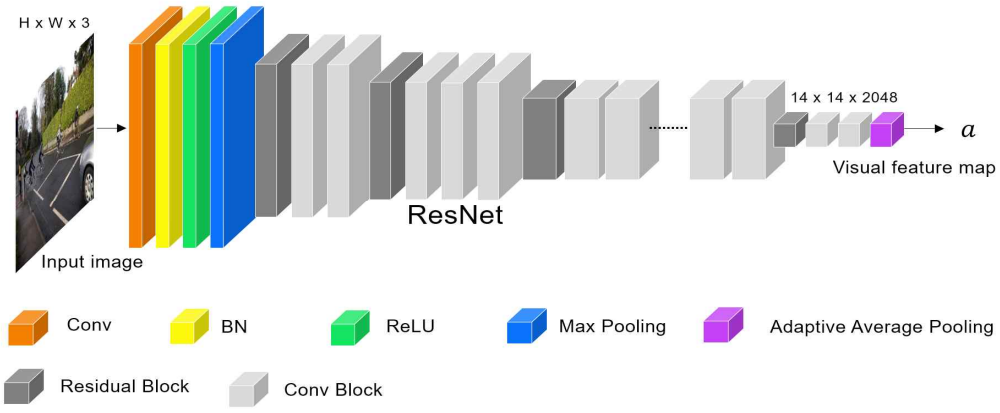

그림 1. 의미적 시각 특징 추출을 위한 인코더
Fig. 1. Encoder for semantic visual feature extraction

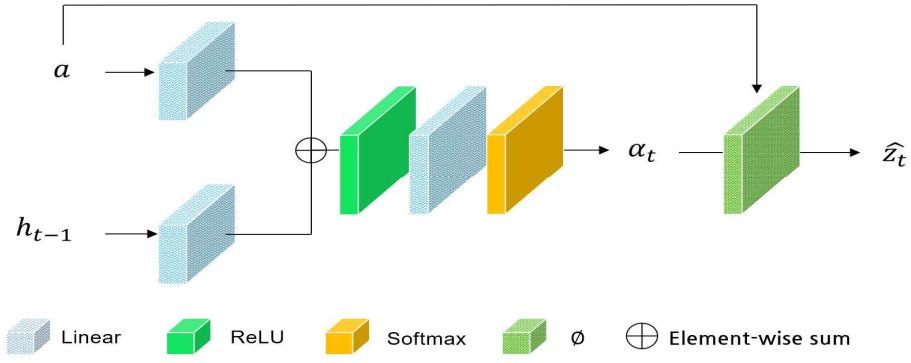

그림 2. 어텐션 네트워크
Fig. 2. Attention network

입력 영상의 정보를 추출해서 차원이 축소된 특징 벡터로 변환하기 때문에 인코딩의 기능과 유사하다고 볼 수 있다. 주로 이미지넷에 사전학습된 ResNet[7] 및 VGG[8] 모델이 인코더로 사용된다. 이때 ResNet 모델을 그대로 사용하는 것이 아니라 맨 마지막 두 개의 레이어를 제거한 네트워크만 사용한다. ResNet 모델은 합성곱(Convolution : Conv) 레이어, 배치 정규화 (Batch Normalization : BN) 레이어, 정류선형유닛(Rectified Linear Unit : ReLU), 최대 풀링(Max pooling) 레이어, 잔차 블록(Residual block), 적응적 평균 풀링(Adaptive average pooling)로 구성된다. 여기서 잔차 블록은 스킵 연결(Skip connection) 레이어와 합성곱 레이어를 병렬로 연결해서 결합하는 방식을 취한다. 즉, 이전 레이어의 특징 맵을 스킵 연결 레이어를 사용해서 출력단으로 정보를 전송하는 역할을 함으로써, 오차가 역전파(Backpropagation)될 때 출력단에서 입력단까지 손실 없이 전파되도록 해준다. 이는 곧 깊은 계층을 쌓을 수 있음을 의미한다. 그리고 적응적 평균 풀링은 입력 영상의 사이즈에 관계없이 맨 마지막 레이어의 출력 맵의 크기를 동일한 사이즈로 만드는 역할을 한다. ResNet 모델에 관한 자세한 사항은 참고문헌을 보시기 바란다[7]. ResNet 기반 인코더는 최종 가로와 세로 길이가 14×14이고 채널 개수가 2048개인 3차원 텐서를 생성한다.

$$a = \{a_1, \cdots, a_L\} \tag{1}$$

여기서 a는 인코더 출력단의 $14\times14\times2048$ 3차원 텐서를 의미한다. 그리고 아래 첨자는 픽셀 인덱스를 의미한다. 예를 들어, a_i는 i번째 픽셀에 위치한

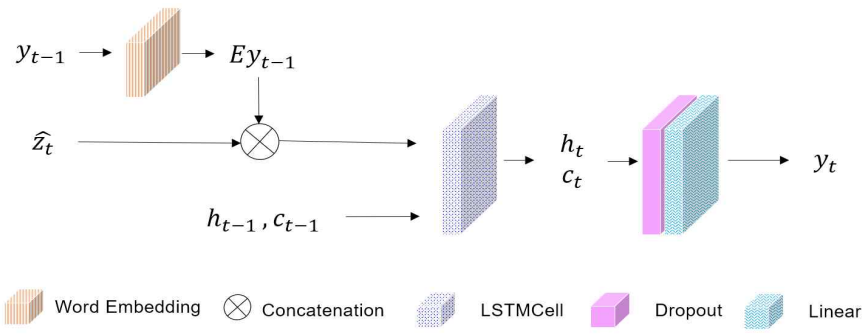

그림 3. 문장 생성을 위한 LSTM
Fig. 3. LSTM for sentence generation

특징 벡터를 의미하고 $1\times1\times2048$의 사이즈를 갖는다.

2.2 디코더

디코더는 인코더에서 추출된 의미적 시각 특징 맵으로부터 문장을 생성하는 역할을 담당한다. 디코더는 크게 어텐션(Attention) 네트워크와 장단기 메모리(Long Short-Term Memory : LSTM)[9]으로 구성된다. 첫째, 어텐션 네트워크는 인코더에서 추출된 시각 특징 맵에서 공간적으로 어떤 영역이 더 중요한지를 판별하는 역할을 한다. 디코더는 시각 특징 맵(a)으로부터 다음에 생성될 단어를 예측한다. 이때, 시각 특징 맵에서 공간적으로 어떤 특징 벡터 a_i가 다음 단어를 예측할 때 가장 큰 영향을 줄지를 판단한다. 즉, 어텐션 네트워크는 시각 특징 맵(a)에 대응하는 가중치 맵을 추정한다고 볼 수 있다. 그림 2는 어텐션 네트워크의 구조이다. 그림에서 보듯이, 어텐션 네트워크는 입력 시각 특징 맵(a_i)과 장단기 메모리에서 출력된 $t-1$번째 은닉 벡터(h_{t-1})로부터 선형(Linear) 레이어와 연결(Concatenation) 레이어 등을 거쳐 가중치 맵인 α_t를 예측한다. 그리고 $\varnothing$함수를 거쳐 최종 가중치가 반영된 시각 특징 맵인 $\hat{z}_t$를 출력한다. 여기서 아래 첨자 t는 장단기 메모리의 타임 스텝을 의미한다. $\hat{z}_t$는 α_t와 a_i를 사용해서 아래와 같이 계산된다.

$$\hat{z}_t = \varnothing(\{a_i\},\{\alpha_i \tag{2}$$

여기서 $\varnothing$는 i번째 픽셀의 시각 특징 벡터 a_i와 가중치 $\alpha_{t,i}$를 가중치 평균을 내는 함수이다. 즉, t번째 단어를 예측할 때, 공간적으로 가장 중요한 시각 특징 벡터를 뽑아내는 과정으로 볼 수 있다.

둘째, 장단기 메모리는 $t-1$번째에서 추정된 단어로부터 t번째 단어를 생성하는 역할을 한다. 그림 3은 문장 생성을 위한 장단기 메모리 구조를 보여주고 있다. 장단기 메모리의 입력은 $t-1$번째에서 추정된 단어(y_{t-1}), t번째 단어 예측을 위해 공간 어텐션이 반영된 시각 특징 맵($\hat{z}_t$)이다. 물론 시계열 데이터에서 시간에 따라 장기적으로 기억해야 될 정보를 간직하고 문맥 의존성을 반영할 수 있는 메모리 셀(c_{t-1})과 은닉 벡터(h_{t-1})를 입력으로 받는다. 그림에서 E는 임베딩(Embedding) 레이어로써, 입력 단어를 의미적 단어 벡터(Semantic word Vector)로 변환해준다. 실제 E는 행렬로 구현되고, 행렬의 초기 값은 사전 학습된 워드-투-벡터(Word2Vec)[10]가 널리 사용된다.

III. 제안한 폭우 영상 생성 모델과 의미적 시각 특징 정합 기반의 이미지 캡셔닝

2절에서 소개된 기존의 이미지 캡셔닝 기술은 청명한 날씨에서 촬영된 고품질의 영상이나 실내에서 촬영된 선명한 영상을 대상으로 우수한 성능을 획득할 수 있다. 하지만, 악천후와 같은 야외 환경에서 촬영된 저화질의 영상에서는 문장 생성에 실패할 수 있다. 특히 폭우 환경에서는 빗줄기 축적 현

상과 빗줄기로 인해 현저한 가시성 저하가 발생된다. 이러한 저화질의 영상에서는 객체의 특징을 추출하기 어렵다. 이로 인해, 단어에 대응하는 객체를 찾을 수 없으므로 이미지 캡셔닝의 성능은 낮아질 수밖에 없다. 따라서 이러한 문제를 해결하기 위해서는 저화질의 영상을 객체를 식별할 수 있을 정도의 영상 품질로 변환하는 과정과 초기 복원 영상으로부터 시각 특징의 의미적 표현력을 제고하는 과정이 필요하다.

3.1 폭우 영상 생성 모델

먼저 폭우 환경에서 폭우 영상(Heavy Rain)이 생성되는 과정을 소개하고자 한다. 폭우 환경에서는 빗줄기 패턴과 빗줄기 축적 효과 때문에 촬영 영상이 전반적으로 뿌옇게 되어 객체의 가시성이 현저히 저하되는 현상이 발생한다. 이러한 폭우 환경에서 촬영된 빗줄기 영상 생성 과정은 다음과 같이 물리적으로 모델링 될 수 있다[11].

$$I = T \odot (J + \sum_{i}^{n} S_i) + (1 - T) \odot A \quad (3)$$

여기서 I는 폭우 영상을 의미하고 J는 빗줄기가 없는 선명한 원본 영상을 의미한다. 그리고 S_i는 i번째 빗줄기 패턴만을 포함한 레이어를 의미하며 T와 A는 각각 투과율 맵(Transmission Map)과 산란광(Airlight)에 해당한다. 그리고 $\odot$는 원소 별 곱을 의미한다.

식 (3)은 이해를 돕기 위해 다음과 같이 두 단계로 다시 작성할 수 있다.

$$R = J + \sum_{i}^{n} S_i \quad (4)$$

$$I = T \odot R + (1 - T) \odot A \quad (5)$$

식 (4)는 빗줄기 패턴이 형성되는 과정을 모델링한 것으로써, 원본 영상 J에 n개의 빗줄기 레이어를 더함으로써 빗줄기 영상(Rain Streak) R을 만든다. 참고로 S_i는 가우시안 잡음을 생성하고 블러링 효과를 적용해서 빗줄기 패턴을 만든다.

식 (5)는 기존의 헤이즈 모델로써 빗줄기 축적 효과까지 반영한 모델에 해당한다. 그리고 식 (5)에서 투과율 맵은 깊이 맵(Depth Map)으로부터 계산되며 산란광은 밝은 영역에서의 픽셀 값으로 생성된다.

그림 4는 식 (5)를 사용해서 생성된 폭우 영상의 예시를 보여 준다. 그림 4(a)는 원본 영상 J에 해당하고 그림 4(b)는 빗줄기 레이어 S_i가 추가된 결과이다. 그림에서 빗줄기 패턴이 형성된 것을 볼 수 있다. 그리고 그림 4(c)는 깊이 맵에 해당하고 그림 4(d)는 식 (5)의 빗줄기 축적 효과까지 반영한 최종 폭우 영상 I에 해당한다.

그림에서 빗줄기 축적 효과로 인해 마치 헤이즈가 발생한 것처럼 가시성이 현저히 떨어지는 것을 볼 수 있다. 특히 먼 거리에 있는 객체일수록 가시성이 떨어지는 것을 볼 수 있다.

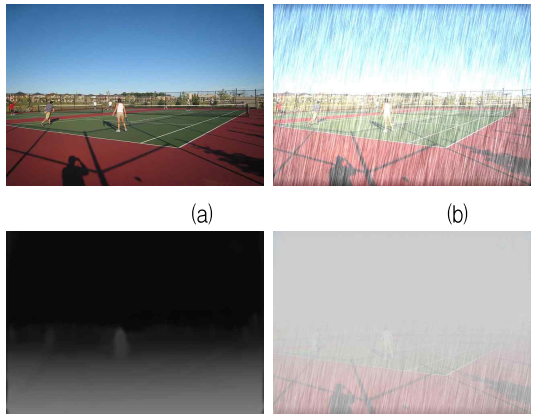

그림 4. 폭우 영상 생성 모델; (a) 원본 영상, (b) 빗줄기 영상, (c) 깊이 맵, (d) 폭우 영상

Fig. 4. Heavy rain image generation model ; (a) original image, (b) rain streak image, (c) depth map, (d) heavy rain image

3.2 폭우 영상 복원

폭우 영상에서 빗줄기 축적과 빗줄기 패턴을 제거하여 선명한 영상을 복구하는 과정은 식 (3)으로부터 유도될 수 있다.

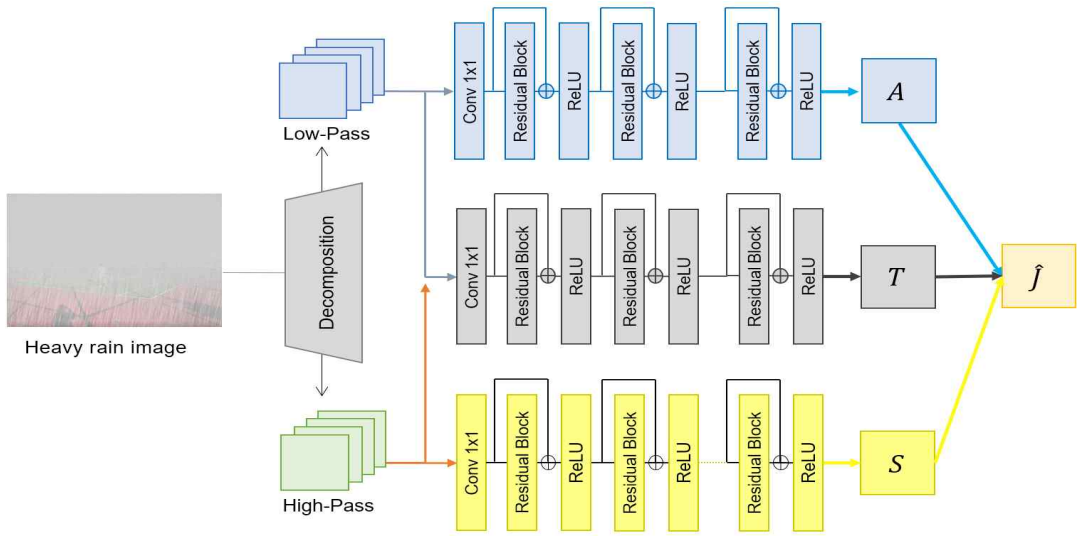

그림 5. 폭우 영상 복원 네트워크
Fig. 5. Heavy rain image restoration network

$$\hat{J} = \frac{I - (1 - T) \odot A}{T} - \sum_{i}^{n} S_i \qquad (6)$$

식 (6)에서 $\hat{J}$는 추정된 복원 영상이다. $\hat{J}$를 추정하기 위해서는 주어진 입력 폭우 영상 I로부터 S, T, A를 추정해야 된다. 이를 위해 본 연구에서는 ResNet 모델을 사용해서 복원하고자 한다.

그림 5는 폭우 영상 복원을 위한 딥 러닝 구조를 보여 주고 있다. 입력 폭우 영상 I에서 영상 분해를 통해서 저주파와 고주파 레이어로 분해한 후, ResNet 서브네트워크를 통해 S, T, A를 각각 추정한다. 영상 분해는 가이드 영상 필터링(Guided Image Filtering)[12]과 같은 적응적인 저주파 필터링을 사용하면 된다. A는 산랑광에 해당하므로 저주파 특성을 지니고 있고 S는 빗줄기 패턴만을 포함하고 있으므로 고주파 특성을 지니고 있다. 따라서 S는 I의 고주파 레이어에서 추정되고 A는 I의 저주파 레이어에서 추정된다. 그리고 T는 투과율 맵이므로 저주파와 고주파 레이어를 둘 다 사용해서 추정한다. S, T, A가 추정되면 최종 식 (6)을 사용해서 복원 영상 $\hat{J}$를 구할 수 있다. 참고로 그림 5에서 S, T, A를 추정하기 위해 3개의 ResNet 네트워크가 사용되었고 각 네트워크의 손실함수는 평균제곱오차(Mean Squared Error)로 정의되었다.

3.3 의미적 시각 특징 정합

그림 1에 제시된 기존의 이미지 캡셔닝의 인코더는 청명한 날씨에 촬영된 영상에 대해서는 의미적 시각 특징을 추출할 수 있다. 하지만 폭우 영상에서는 현저한 가시성 저하로 객체에서 유용한 시각 특징을 추출하기 힘들다. 물론 앞 절에서 언급한 폭우 영상 복원 과정을 적용하면 의미적 시각 특징을 개선할 수 있지만 화질 개선의 여지가 남아있다. 이를 극복하기 위해, 추정된 복원 영상 $\hat{J}$에서 추출된 시각 특징 벡터와 원본 영상 J에서 추출된 시각 특징 벡터를 정합하고자 한다.

그림 6은 본 연구에서는 제안한 의미적 시각 특징 정합을 위해 개선된 인코더를 보여 주고 있다. 그림에서 보듯이, $\hat{J}$와 J에서 각각 시각 특징을 추출하기 위해 그림 1에서 제시된 2개의 동일한 ResNet 모델을 사용했다. ResNet에서 출력된 두 특징 간의 오차를 최소화함으로써, $\hat{J}$의 시각 특징의 의미적 표현을 개선할 수 있다. 즉, 청명한 날씨에서 촬영된 J의 시각 특징에 가깝도록 변환할 수 있다. 의미적 시각 특징 정합을 위해 l_1-놈(Norm)을 손실함수로 사용했다.

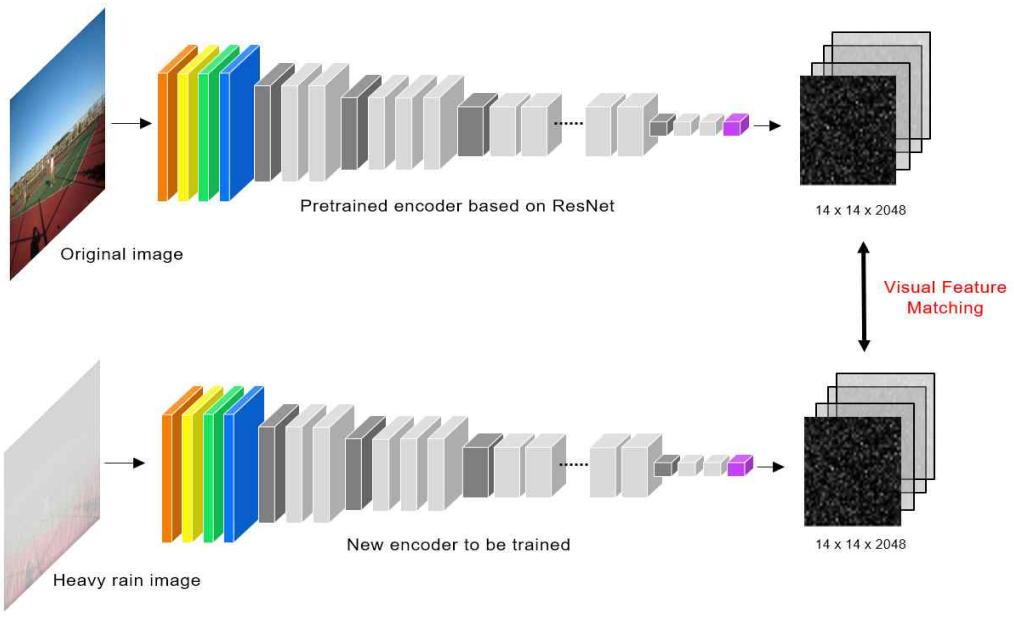

그림 6. 제안한 시각 특징 정합 네트워크
Fig. 6. Proposed network for visual feature matching

IV. 실험 및 결과

학습 데이터는 이미지 캡셔닝에서 널리 사용되는 MSCOCO2014[13]를 사용했다. 약 12만 장의 영상으로 구성된 데이터 집합을 훈련 집합 11만 장, 검증 집합 5천 장, 테스트 집합 5천 장으로 나누어 사용했다. 각 영상 당 5개의 정답 문장이 제공된다. 본 연구에서는 폭우 환경을 대상으로 실험을 하므로 MSCOCO의 선명한 원본 영상에서 폭우 영상을 생성해야 한다. 이를 위해, 3.1절에서 제시된 폭우 영상 생성 모델을 적용해서 폭우 영상을 만들었다. 한편, 본 연구에서는 PyTorch 프레임워크를 사용했고 윈도우 운영체제에서 실험하였다.

4.1 정성적 평가

제안한 이미지 캡셔닝 기법을 평가하기 위해, 2절에 소개된 기존의 어텐션 기반 이미지 캡셔닝[6]과 제안한 폭우 영상 복원과 의미적 시각 특징 정합 기반의 이미지 캡셔닝 기법을 비교하였다. 그림 7은 이미지 캡셔닝의 결과로써, 원본 영상, 폭우 영상, 복원 영상, 그리고 생성된 문장 결과를 보여 주고 있다. 먼저 'Bike' 영상의 경우, 기존의 이미지 캡셔닝 기법은 원본 영상의 내용과 전혀 무관한 "욕조와 샤워 커튼이 있는 욕실"이라는 잘못된 문장을 생성한 것을 볼 수 있다. 반면에 의미적 시각 특징 정합과 폭우 영상 복원 과정을 모두 적용한 제안한 기법은 "거리에서 자전거를 타고 있는 한 무리의 사람들"이라는 정확한 문장을 생성한 것을 볼 수 있다. 이는 폭우 환경에서 정확한 이미지 캡셔닝 기술을 달성하기 위해서는 폭우 영상의 화질 개선과 시각 특징의 의미적 표현력을 제고해야 함을 말해준다. 그리고 제안한 기법이 폭우 환경에서의 이미지 캡셔닝의 성능을 개선할 수 있다는 것을 확증해 준다. 다른 입력 영상에 대해서도 유사한 결과가 나타난다. 예를 들어, 'Tennis' 영상에 대해서도 기존의 이미지 캡셔닝 기법은 "침대 위에 고양이 사진"이라는 전혀 이상한 문장을 생성한 것을 볼 수 있다. 반면 제안한 기법은 "테니스 코트에서 테니스 라켓을 들고 있는 남자"라는 비교적 정확한 문장을 생성한 것을 볼 수 있다. 'Beach' 영상에 대해서도 기존의 이미지 캡셔닝 기술은 "눈이 덮힌 필드 위에 서 있는 사람들의 그룹"이라는 전혀 엉뚱한 문장을 생성한 것을 볼 수 있다. 이는 빗줄기

축적 현상 때문에 폭우 영상의 전반적인 색상이 하얗게 변했기 때문으로 분석된다. 반면 제안한 기법에서 “들판에서 연을 날리는 한 무리의 사람들”이라는 상대적으로 정확한 문장을 생성한 것을 볼 수 있다.

참고로 폭우 영상의 복원은 이미지 캡셔닝의 최종 목적이 아니다. 폭우 영상을 복구하는 목적은 이미지 캡셔닝의 인코더에 추출된 시각 특징의 의미적 표현력을 개선하기 위함이다. 그림 7에서 비록 복원 영상의 화질이 선명한 영상은 아닐지라도 인코더의 시각 특징 추출 능력을 향상할 수 있다.

4.2 정량적 평가

이미지 캡셔닝에서 주로 사용되는 평가 방법으로는 BLEU(Bilingual Evaluation Understudy)와 METEOR(Metric for Evaluation of Translation with Explicit Ordering)[2][3]가 있다. BLEU는 기계 번역에서 원본 문장과 번역 문장이 얼마나 유사한지를 판단하기 위해 사용되는 척도이다. BLEU도 n-gram에 따라 문장의 유사도를 평가할 수 있다. 여기서 n-gram이란 문장의 유사도를 계산할 때, 몇 개의 단어를 묶어서 평가할지를 정하는 기준을 말한다. 만약 n=1이면, 한 단어씩 유사도를 평가하고 만약 n=2면 두 개의 단어를 한 묶음으로 생각하여 두 문장의 겹친 정도를 평가한다. 이와 유사하게 METEOR도 문장의 조화 평균을 사용하여 원본 문장과 생성된 문장과의 유사도를 계산한다.

표 1은 기존 이미지 캡셔닝 기법과 제안한 기법에 대한 정량적 평가 결과를 보여 준다. 표에서 볼 수 있듯이, 기존의 이미지 캡셔닝 기법을 사용하여 폭우 영상에 대해 생성한 문장의 BLEU와 METEOR 점수는 상대적으로 낮은 것을 볼 수 있다. 이는 폭우 영상의 가시성 저하와 폭우 축적 효과로 폭우 영상의 색상 톤의 왜곡에 기인한 것으로 분석된다. 반면 제안한 기법을 적용했을 때 BLEU와 METEOR 점수는 기존의 이미지 캡셔닝 기법보다 각 평가 항목에서 약 7% 이상의 성능을 개선한 것을 볼 수 있다. 이는 물리적 기반으로 폭우 모델링을 통해 폭우 영상의 화질을 개선한 후, 시각 특징 정합 과정을 통해 의미적 표현력을 향상했기 때문이다. 따라서 정량적 수치 비교를 통해서도 제안한 물리적 폭우 모델링과 의미적 시각 특징 정합 기반의 인코더가 기존의 이미지 캡셔닝 기술을 상당히 개선할 수 있음을 확인할 수 있다.

표 1. 이미지 캡셔닝의 정량적 평가
Table 1. Quantitative evaluation for image captioning

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
기존의 방법[6]	52.84	32.40	19.97	12.87	29.39
제안한 기법	60.08	42.27	30.20	21.52	37.44

V. 결 론

본 연구에서는 폭우 환경에서도 이미지 캡셔닝의 성능을 개선할 수 있는 새로운 인코더를 제안하였다. 폭우 환경에서 촬영된 영상의 가시성 저하 문제를 해결하기 위해, 먼저 폭우 영상 생성 모델을 기반으로 훈련 데이터를 구축하고 산란광, 투과율, 빗줄기 레이어를 ResNet 네트워크를 학습하여 추정하였다. 그리고 폭우 영상 생성 모델의 역과정을 계산함으로써, 화질이 개선된 복원 영상을 획득하였다. 또한 복원 영상의 시각 특징 벡터의 의미적 표현력을 제고하기 위해, 원본 영상의 시각 특징 벡터와 복원 영상의 시각 특징 벡터를 각각 ResNet 네트워크를 통해 정합하는 과정을 수행하였다. 이를 통해, 제안한 이미지 캡셔닝 기법은 기존의 이미지 캡셔닝 기법보다 BLEU와 METEOR 성능 지표에서 약 7% 이상을 개선할 수 있었다.

향후 연구로써는 한글 전용 이미지 캡셔닝 기술을 개발하고자 한다. 본 연구에서 개발된 이미지 캡셔닝 기법은 입력 영상으로부터 영어 문장만을 생성할 수 있다. 하지만 국내 환경에 적용하기 위해서는 한글 임베딩을 통한 한글 문장 생성 기술을 개발할 필요가 있다. 그리고 이 연구에서 실험된 데이터는 합성 영상이다. 실제 폭우 영상과는 다를 수 있다. 따라서 이러한 격차를 극복하기 위해 실제 폭우 영상 데이터를 학습하여 이미지 캡셔닝 고도화 작업을 추진하고자 한다.

Original 'Bike' image　　Heavy rain image　　Restored image

Ground truth	a group of people riding bikes down a street
Conventional image captioning [6]	a bathroom with a tub and a shower curtain
Proposed image captioning	a group of people riding bikes down a street

(a)

Original 'Beach' image　　Heavy rain image　　Restored image

Ground truth	a large group of people flying kites on a beach
Conventional image captioning [6]	a group of people standing on top of a snow covered field
Proposed image captioning	a large group of people flying kites in a field

(b)

Original 'Fire Hydrant' image　　Heavy rain image　　Restored image

Ground truth	a group of people standing around a fire hydrant
Conventional image captioning [6]	a group of people standing on a snow covered slope
Proposed image captioning	a group of people standing next to each other

(c)

Original ‘Fruit Stand’ image　　Heavy rain image　　Restored image

Ground truth	a group of people standing around a fruit stand
Conventional image captioning [6]	a group of people standing on top of a snow covered slope
Proposed image captioning	a group of people standing around a bunch of bananas

(d)

Original ‘Park Bench’ image　　Heavy rain image　　Restored image

Ground truth	a group of people sitting on a park bench
Conventional image captioning [6]	a group of people standing in front of a building
Proposed image captioning	a group of people sitting on a bench

(e)

Original ‘Tennis’ image　　Heavy rain image　　Restored image

Ground truth	a couple of people playing a game of tennis
Conventional image captioning [6]	a picture of a cat on a bed
Proposed image captioning	a man holding a tennis racquet on a tennis court

(f)

그림 7. 이미지 캡셔닝 결과 ; (a) ‘Bike’ 영상, (b) ‘Beach’ 영상, (c) Fire Hydrant’ 영상, (d) ‘Fruit Stand’ 영상, (e) ‘Park Bench’ 영상, (f) ‘Tennis’ 영상

Fig. 7. Results of image captioning ; (a) ‘Bike’ image, (b) ‘Beach’ image, (c) ‘Fire Hydrant’ image, (d) ‘Fruit Stand’ image, (e) ‘Park Bench’ image, (f) ‘Tennis’ image

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, No. 4, pp. 652-663, Apr. 2017.

[2] L. Chen et al., "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning", in Proc. IEEE International Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 6298-6306, Jul. 2017.

[3] J. Mao et al., "Deep captioning with multimodal recurrent neural networks", in Proc. International Conference on Learning Representations, arXiv:1412.6632 [cs.CV], Jun. 2015.

[4] C.-H. Son and X.-P. Zhang, "Rain detection and removal via shrinkage-based sparse coding and learned rain dictionary", Journal of Imaging Science and Technology, Vol. 64, No. 3, pp. 30501-1-30501-17(17), May. 2020.

[5] J.-C Kim and C.-H. Son, "Structure-aware residual network for rain streak removal", The Journal of Korean Institute of Information Technology, Vol. 18, No. 10, pp. 87-100, Oct. 2020.

[6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. "Show, attend and tell: Neural image caption generation with visual attention", In International Conference on Machine Learning, Jun. 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in Proc. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, pp. 770-778, Jun. 2016.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", arXiv: 1409.1556 [cs. CV], Sep. 2014.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory", Neural Computation, Vol. 9, No. 8, pp. 1735-1780, Dec. 1997.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality", arXiv:1310.4546 [cs.CL], Oct. 2013.

[11] W. Yang, R. T. Tan, S. Wang, Y. Fang, and J. LiuSingle, "Image deraining: From model-based to data-driven and beyond," arXiv:1912.07150 [eess.IV], Dec. 2019.

[12] K. He, J. Sun, and X. Tang, "Guided image filtering", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35, No. 6, pp. 1397-1409, Jun. 2013.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft coco: Common objects in context", in Proc. European Conference on Computer Vision, arXiv:1405.0312 [cs.CV], May. 2014.

저자소개

여 풍 휘 (Pung-Hwi Ye)

2018년 3월 ~ 현재 : 군산대학교 소프트웨어융합공학과 학사과정

관심분야 : 컴퓨터 비전, 영상처리, 딥 러닝

손 창 환 (Chang-Hwan Son)

2002년 2월 : 경북대학교 전자전기공학부(공학사)

2004년 2월 : 경북대학교 전자공학과(공학석사)

2008년 8월 : 경북대학교 전자공학과(공학박사)

2017년 4월 ~ 현재 : 군산대학교 소프트웨어융합공학과 부교수

관심분야 : 컴퓨터 비전, 영상처리, 딥 러닝, 기계학습, 색 재현