ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2020 June: 26(2): 79~104 http://dx.doi.org/10.13088/jiis.2020.26.2.079

전문성 이식을 통한 딥러닝 기반 전문 이미지 해석 방법론*

김태진
국민대학교 비즈니스IT 전문대학원
(*jeung722@kookmin.ac.kr*)

김남규
국민대학교 비즈니스IT 전문대학원
(*ngkim@kookmin.ac.kr*)

최근 텍스트와 이미지 딥러닝 기술의 괄목할만한 발전에 힘입어, 두 분야의 접점에 해당하는 이미지 캡셔닝에 대한 관심이 급증하고 있다. 이미지 캡셔닝은 주어진 이미지에 대한 캡션을 자동으로 생성하는 기술로, 이미지 이해와 텍스트 생성을 동시에 다룬다. 다양한 활용 가능성 덕분에 인공지능의 핵심 연구 분야 중 하나로 자리매김하고 있으며, 성능을 다양한 측면에서 향상시키고자 하는 시도가 꾸준히 이루어지고 있다.

하지만 이처럼 이미지 캡셔닝의 성능을 고도화하기 위한 최근의 많은 노력에도 불구하고, 이미지를 일반인이 아닌 분야별 전문가의 시각에서 해석하기 위한 연구는 찾아보기 어렵다. 동일한 이미지에 대해서도 이미지를 접한 사람의 전문 분야에 따라 관심을 갖고 주목하는 부분이 상이할 뿐 아니라, 전문성의 수준에 따라 이를 해석하고 표현하는 방식도 다르다. 이에 본 연구에서는 전문가의 전문성을 활용하여 이미지에 대해 해당 분야에 특화된 캡션을 생성하기 위한 방안을 제안한다.

구체적으로 제안 방법론은 방대한 양의 일반 데이터에 대해 사전 학습을 수행한 후, 소량의 전문 데이터에 대한 전이 학습을 통해 해당 분야의 전문성을 이식한다. 또한 본 연구에서는 이 과정에서 발생하게 되는 관찰간 간섭 문제를 해결하기 위해 '특성 독립 전이 학습' 방안을 제안한다. 제안 방법론의 실현 가능성을 파악하기 위해 MSCOCO의 이미지-캡션 데이터 셋을 활용하여 사전 학습을 수행하고, 미술 치료사의 자문을 토대로 생성한 '이미지-전문 캡션' 데이터를 활용하여 전문성을 이식하는 실험을 수행하였다. 실험 결과 일반 데이터에 대한 학습을 통해 생성된 캡션은 전문적 해석과 무관한 내용을 다수 포함하는 것과 달리, 제안 방법론에 따라 생성된 캡션은 이식된 전문성 관점에서의 캡션을 생성함을 확인하였다.

본 연구는 전문 이미지 해석이라는 새로운 연구 목표를 제안하였고, 이를 위해 전이 학습의 새로운 활용 방안과 특정 도메인에 특화된 캡션을 생성하는 방법을 제시하였다.

주제어 : 딥러닝, 전문성 이식, 전이 학습, 이미지 캡셔닝, 인공지능

논문접수일 : 2020년 5월 12일 논문수정일 : 2020년 6월 4일 게재확정일 : 2020년 6월 20일
원고유형 : 일반논문 교신저자 : 김남규

1. 서론

최근 다양한 분야에서 데이터 기반 의사결정 문제를 더욱 빠르고 정확하게 해결하기 위한 방안으로 딥러닝(Deep Learning)에 대한 관심이 급증하고 있다. 딥러닝은 인간의 신경계와 유사한 구조를 가진 기계 학습 알고리즘의 일종으로, 데이터에 내재된 유의미한 특성(Feature)을 자체적

* 본 논문은 과학기술정보통신부 및 정보통신산업진흥원의 '고성능 컴퓨팅 지원' 사업으로부터 지원받아 수행하였음.

으로 발견하고, 이렇게 도출된 특성들로부터 목적 레이블(Target Label)을 추론하기 위한 학습을 수행한다는 점에서 기존의 기계 학습 알고리즘과 차이가 있다. 딥러닝은 구매 예측, 고객 이탈 예측, 추천시스템, 그리고 기업의 부도 예측 등의 분야에서 기존의 예측 기법인 인공신경망, 의사결정나무, 로지스틱 회귀, 그리고 SVM 등을 빠르게 대체하고 있으며(Caigny et al, 2019; Chen et al, 2019; Feng et al, 2019; Pang et al, 2019), 최근에는 가짜 뉴스 식별, 재난 탐지, 그리고 온라인 범죄 예측 등 다양한 사회적 문제 해결에 딥러닝이 활용되고 있다(Liu et al, 2016; Sanjiban et al, 2017; Yang et al, 2018).

분석 데이터 측면에서는 텍스트 데이터와 이미지 데이터에 대한 딥러닝 연구가 가장 활발하게 이루어지고 있다. 텍스트 딥러닝 연구는 텍스트의 문맥에 대한 학습을 통해 텍스트에 담긴 고유의 특성을 추출하여 이를 벡터(Vector)로 표현하기 위한 임베딩(Embedding) 기법을 출발점으로, 문서 분류, 감성 분석 등 다양한 분야에서 진행되고 있다(Tomas et al, 2013; Hochreiter et al, 1997; Devlin et al, 2018; Peters et al, 2018; Yang et al, 2019). 한편 이미지 데이터에 대한 딥러닝 연구는 CNN(Convolutional Neural Network) 모델을 기반으로 다양한 후속 연구들이 활발하게 이루어지고 있다. CNN 계열의 다양한 모델들은 이미지 분류와 객체 검출 등 이미지 분석과 관련된 전반적인 분야에서 우수한 성능을 보이고 있으며, 최근에도 'ImageNet' 이나 'MSCOCO' 등 이미지 딥러닝 대회를 통해 우수한 성능의 이미지 딥러닝 모델들이 꾸준히 공개되고 있다(Forrest et al, 2016; Huang et al, 2017; Ren et al, 2015; Alex et al, 2012; Christain et al, 2015; He et al, 2016).

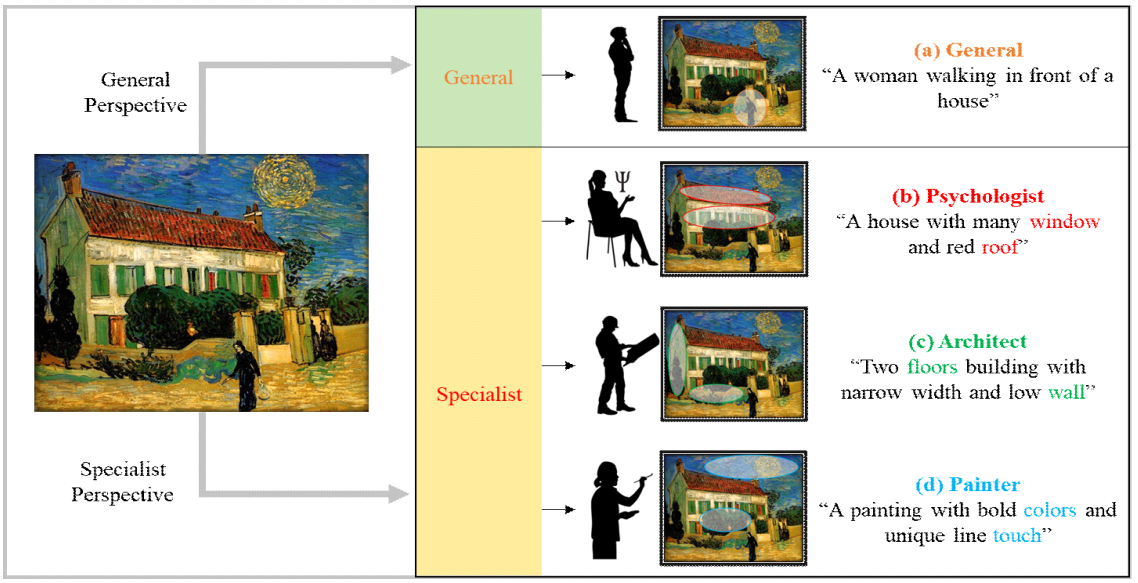
최근에는 이미지 딥러닝과 텍스트 딥러닝 기술의 괄목할 만한 발전에 힘입어, 두 분야의 접점에 해당하는 이미지 캡셔닝(Image Captioning)의 활용 및 기술에 대한 관심이 급증하고 있다. 이미지 캡셔닝은 입력 이미지를 이해하고 그에 적합한 캡션을 출력으로 생성하는 기술이며, 이미지 인코딩과 텍스트 생성을 동시에 다룬다(Ryan et al, 2014). 이미지 캡셔닝은 기본적으로 이미지 인덱싱 및 검색에 사용될 수 있으며, 의학, 심리학, 교육, 그리고 소셜 미디어 등 다양한

〈Figure 1〉 Seeing AI – Application of Image Captioning

분야에서 활용될 수 있다. 이미지 캡셔닝의 가장 널리 알려진 응용으로 시각 장애인의 눈을 대신하여 카메라에 보이는 사람, 제품, 장면 등을 음성으로 설명해 주는 모바일 앱인 Microsoft의 'Seeing AI'를 들 수 있다. <Figure 1>은 'Seeing AI'의 실제 시연 장면으로, 스마트폰 카메라에 포착된 이미지에서 사람과 사물을 비교적 정확하게 파악하여 캡션을 생성했음을 확인할 수 있다.

이미지 캡셔닝의 성능을 다양한 측면에서 향상시키고자 하는 시도가 꾸준히 이루어지고 있다. 특히 단어를 생성할 때마다 이미지의 특정 영역에 집중하는 어텐션(Attention) 기법을 이미지 캡셔닝에 적용하면서 성능이 한 단계 도약했다(Ashnish et al, 2017). 최근에는 이미지를 단순히 정확하게 기술하는 것을 넘어, 이미지에 내재된 정보를 더욱 세련되게 전달할 수 있는 고급 캡션을 생성하기 위한 연구가 이루어지고 있다. 하지만 이처럼 고급 캡션을 생성하기 위한 최근의 많은 노력에도 불구하고 이미지를 일반인이 아닌 분야별 전문가(Domain Experts)의 시각에서 해석하기 위한 연구는 찾아보기 어렵다. 동일한 이미지에 대해서도 이미지를 접한 사람의 전문 분야에 따라 관심을 갖고 주목하는 부분이 상이할 뿐 아니라, 전문성의 수준에 따라 이를 해석하고 표현하는 방식도 다르게 나타나게 된다. 일반인의 경우 전체적이고 일반적인 시각, 즉 이미지의 구성 객체와 그 관계를 식별하는 관점에서 이미지를 인식하는 경향이 있으며, 전문가의 경우 전문성을 바탕으로 주어진 이미지를 해석하기 위해 필요한 요소에 집중하여 이미지를 인식하는 경향이 있다. 이처럼 동일한 이미지라 할지라도 바라보는 사람의 관점에 따라 중요하게 인식하는 내용이 상이하게 나타나며, 이는 <Figure 2>를 통해 설명할 수 있다.

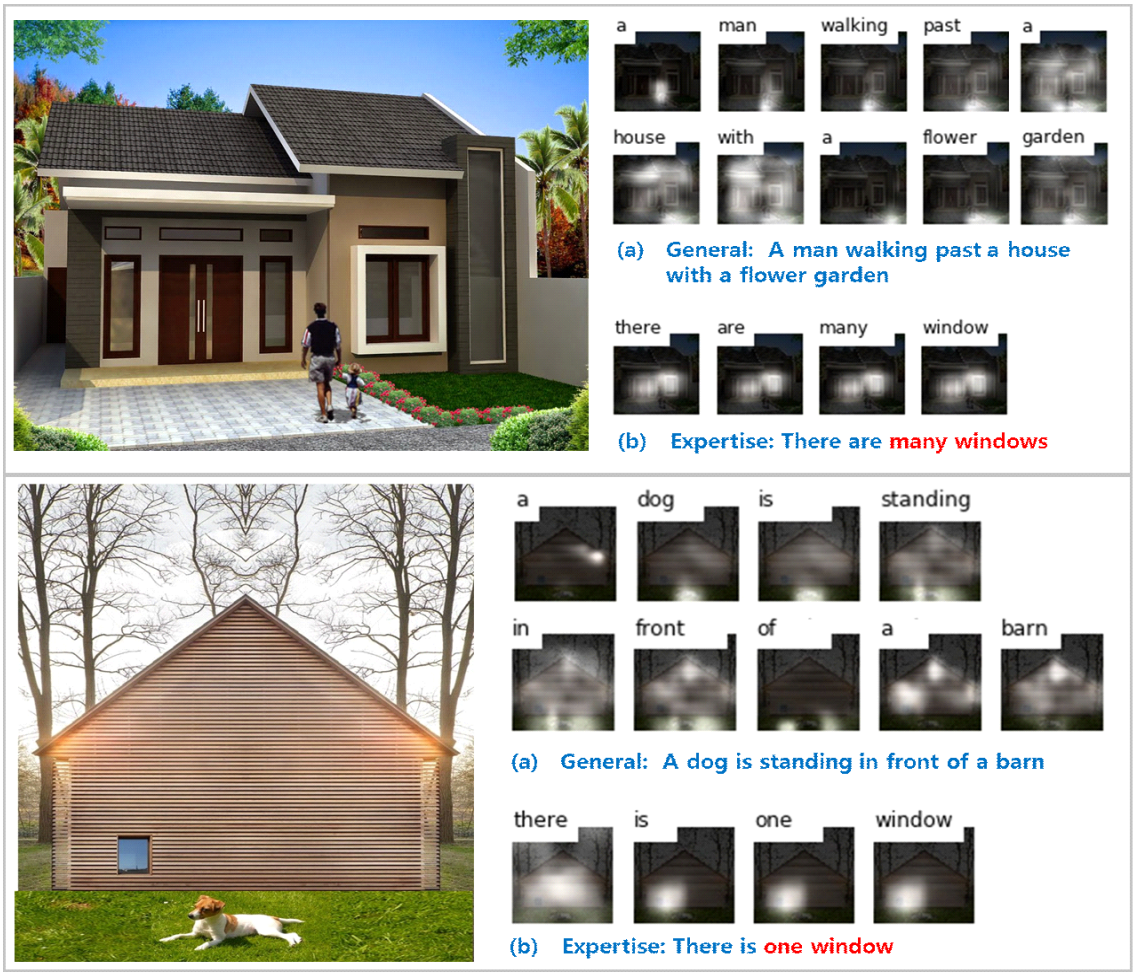
<Figure 2(a)>는 주어진 그림에 대해 일반인이 캡션을 부여한 예이며, <Figure 2(b)>와 <Figure



〈Figure 2〉 Different Captions with Various Expertise

2(c)> 그리고 <Figure 2(d)>는 각각 미술 치료사와 건축가 그리고 화가가 각자의 전문성을 바탕으로 캡션을 부여한 예이다. 예를 들어 미술 치료사의 경우 지붕과 창문에 집중하여 캡션을 부여했는데, 실제로 미술 치료에서 지붕은 생활의 공상 영역을 상징하며, 창문은 환경과의 접촉을 나타내는 중요한 단서이다. 한편 건축가는 건물의 층수와 넓이 및 벽의 높이에 주목하고 있다. 이처럼 동일한 이미지라 할지라도 바라보는 사람의 관심 및 전문성에 따라 인식하는 내용이 상이하게 나타나는 것은 지극히 당연한 현상이므로, 이미지로부터 캡션을 생성하는 이미지 캡셔닝 기법도 이러한 현상을 반드시 반영할 필요가 있다.

이에 본 연구에서는 전문가의 전문성을 활용하여 이미지에 대해 해당 분야에 특화된 전문 캡션을 도출하기 위한 방안을 제안하고자 한다. <Figure 3>은 미술치료 전문가의 관점에서 캡션



〈Figure 3〉 Preliminary Results of Expertized Image Captioning

을 생성한 실제 실험 결과의 일부를 나타내고 있다. 미술치료 전문가는 집 그림을 통해 심리를 해석할 때 문, 창문, 굴뚝 등을 중요한 객체로 생각하며, 특히 창문의 경우 개수, 모양, 크기 등을 중요한 특징으로 다룬다. 예를 들어 어떤 아이가 그린 그림에 창문이 많이 포함된 경우, 이 아이는 외부와 접촉하고자 하는 강한 욕구를 갖고 있는 것으로 해석될 수 있다. 본 실험은 주어진 이미지에 대해 창문의 수에만 집중한 캡션을 생성하는 것을 목적으로 수행되었으며, 각 그림에서 (a) General Caption은 일반적인 관점에서 생성한 캡션을, 그리고 (b) Expertise Caption은 미술 치료 전문가의 관점에서 생성한 캡션을 나타내고 있다. 또한 우측의 작은 이미지들의 집합은 원본 이미지의 어떤 부분이 해당 단어의 생성에 기여했는지를 나타내고 있다. 제시된 그림에서 일반적인 캡션은 ‘man’, ‘house’, ‘flower’, ‘dog’, 그리고 ‘barn’ 등 일반적인 객체의 인식에 초점을 두는 반면 전문적인 캡션은 본 실험의 목적인 창문의 수에 초점을 둔 캡션을 생성함을 확인할 수 있다.

즉, 딥러닝 학습을 위해 이미지/캡션 쌍(Pairs)을 사용하게 되는데, 이 때 일반적인 캡션이 아닌 특정 분야의 전문가에 의해 작성된 캡션을 사용함으로써 전문적인 캡션을 학습할 수 있다. 하지만 이처럼 전문가에 의해 작성된 이미지/캡션 쌍은 그 수가 매우 부족하며 확보도 어려우므로, 기존의 이미지 캡셔닝 기법을 통해 각 분야의 전문적 캡션을 자동으로 생성하기란 현실적으로 불가능하다. 따라서 본 연구에서는 전이 학습, 즉 방대한 양의 일반 데이터에 대해 사전 학습을 수행한 후, 소량의 전문 데이터에 대한 미세 조정을 진행하여 이러한 문제를 해결하고자 한다.

하지만 단순히 전문적 캡션을 사용하여 전이 학습을 수행하는 것은 또 다른 유형의 한계를 야기한다. 미술 치료사가 그림을 해석할 때 문의 유무, 창문의 개수, 지붕의 형태 등을 관찰하는 것과 같이, 전문가가 이미지를 인식할 때 하나의 특성이 아니라 여러 특성을 동시에 관찰하게 된다. 이처럼 여러 관찰, 즉 복수의 특성에 대한 기술이 혼재된 채로 학습이 이루어지면, 관찰간 간섭(Interference)이 발생하여 각 특성 관점의 순수한 학습이 이루어지기 어렵다. 방대한 데이터에 대한 학습에서는 이러한 간섭의 상당 부분이 자체적으로 정화되어 학습 결과에 거의 영향을 미치지 않지만, 소량의 데이터에 대해 학습이 이루어지는 미세 조정의 경우 이러한 간섭이 학습에 미치는 영향이 상대적으로 매우 클 수 있다. 즉 관찰된 여러 특성들이 서로 간섭을 일으켜 각 특성에 대한 학습을 방해할 수 있다. 따라서 본 연구에서는 전문가가 여러 개별 특성을 동시에 관찰하고 그 특성을 종합하여 이미지를 해석하는 과정에 착안하여, 종합적인 해석을 생성하기에 앞서 각 특성에 대한 ‘관찰 캡션’을 분리하여 생성하고, 이를 종합하여 ‘전문 해석’을 도출하는 방안을 제시하고자 한다.

2. 관련 연구

2.1 텍스트와 이미지 딥러닝

텍스트 딥러닝 연구는 텍스트의 문맥에 대한 학습을 통해 텍스트에 담긴 고유의 특성을 추출하여 이를 벡터로 표현하기 위한 임베딩 기법을 주로 다루고 있다. 구체적으로 단어 임베딩(Word Embedding)은 단어 단위의 텍스트를 벡터로 변환하는 과정을 다루며, word2vec(Tomas et

al, 2013), glove(Jeffrey et al, 2014), 그리고 fasttext(Piotr et al, 2016) 등의 모델이 널리 사용되고 있다. 한편 문서 임베딩(Document Embedding)은 문서 내에 존재하는 단어들의 의미를 추론하고, 이를 문서 단위의 벡터로 나타내는 방법을 학습한다(Tomas et al, 2013). 하지만 전통적인 임베딩 모델은 단어의 의미 추론에 단어 주변의 국소적인 문맥만을 참조하므로, 텍스트의 전체적인 문맥과 의미를 충분히 임베딩에 반영하기 어렵다. 이러한 한계를 해결하기 위해 은닉층 노드의 출력을 다음 노드의 입력으로 전달함으로써, 텍스트의 전체 시퀀스(Sequence)를 연속적으로 학습에 사용하는 신경망 모델인 순환 신경망(Recurrent Neural Network)이 제안되었다(Micheal et al, 1990). 하지만 순환 신경망 역시 입력 텍스트의 시퀀스가 길어질수록 과거에 등장한 단어의 정보를 전달하는 신호가 약해진다는 한계를 갖고 있으므로, 이를 극복하기 위해 단어의 상태 정보를 더욱 길게 유지할 수 있는 LSTM(Long Short-Term Memory) 알고리즘이 고안되어 널리 사용되고 있다(Hochreiter et al, 1997). 하지만 LSTM 역시도 장기 의존성 문제를 완벽하게 해결하지 못한다는 한계가 있으며, 이를 해결하기 위해 등장한 것이 어텐션 메커니즘이다(Ashnish et al, 2017). 어텐션은 문장에서 학습에 필요한 중요한 정보에 초점을 맞추는 방식으로, 처리해야 할 정보의 양을 줄여준다는 장점을 갖는다. 최근 많은 연구에서 어텐션 메커니즘이 적용되고 있으며, 이는 그 동안 연구된 딥러닝 기법의 성능을 한 차원 끌어 올리는데 기여하였다.

한편 충분한 양의 학습 데이터 확보가 어렵고 학습에 많은 시간이 소요된다는 딥러닝 알고리즘의 한계를 해결하기 위해, 최근에는 사전 학습 언어 모델(Pre-trained Language Model)을 바탕으로 분야별 추가 학습을 통해 미세 조정(Fine Tuning)을 진행하는 전이 학습(Transfer Learning)에 대한 연구가 활발하게 이루어지고 있다. 대표적인 신경망 기반 사전 학습 언어 모델로는 ELMo(Embeddings from Language Model), BERT(Bidirectional Encoder Representations from Transformer)(Devlin et al, 2018; Peters et al, 2018), 그리고 XLNet(eXtra Long Network)(Yang et al, 2019) 등이 널리 알려져 있다. ELMo는 LSTM을 활용하여 텍스트 시퀀스를 정방향, 역방향의 두 방향으로 학습하는 양방향 학습 언어 모델이다. BERT 역시 양방향 학습 언어 모델이며, ELMo에서 나타나는 신호의 전달 강도 한계를 극복하기 위해 특정 단어에 대해 동일 시퀀스에 존재하는 다른 단어와의 관계를 파악하는 알고리즘인 어텐션 메커니즘 기반의 학습을 수행한다. 최근에 고안된 사전 학습 언어 모델인 XLNet은 전체 텍스트를 부분으로 분할하여 학습을 수행하기 때문에 장문의 텍스트에 대한 학습이 가능하다는 점, 그리고 텍스트의 시퀀스를 무작위로 바꾼 학습을 수행하여 원래 텍스트가 가지고 있던 자연스러운 문맥을 더욱 정확하게 파악할 수 있다는 점으로 인해 최근 텍스트 분석의 다양한 분야에서 많은 관심을 받고 있다.

2.2 이미지 캡셔닝 연구

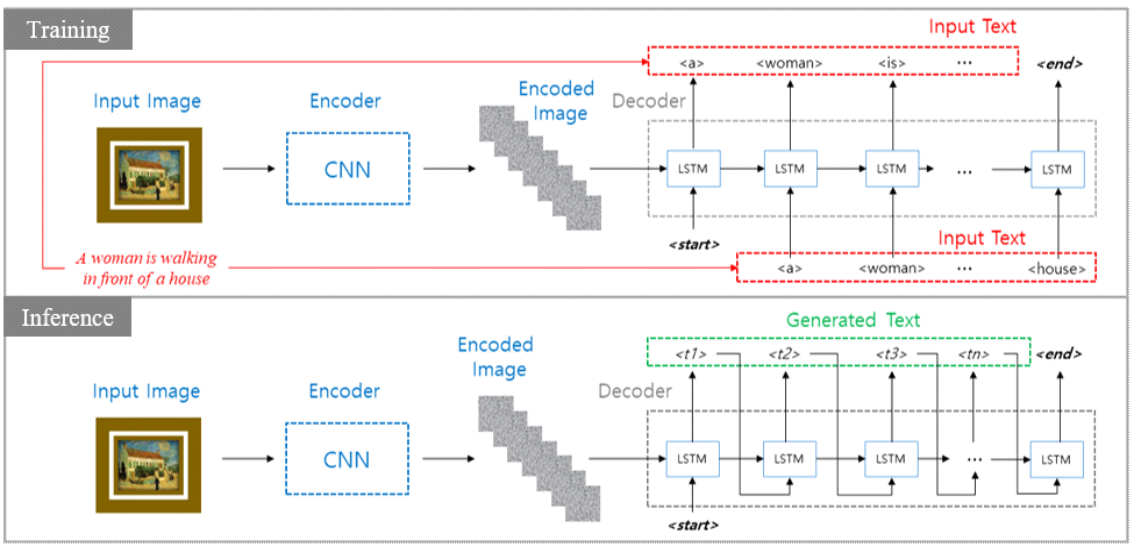
이미지 데이터에 대한 딥러닝 연구는 CNN(Lecun et al, 1989) 모델을 기반으로 다양한 후속 연구들이 활발하게 이루어지고 있다(Forrest et al, 2016; Huang et al, 2017; Ren et al, 2015). CNN 계열의 다양한 모델들은 이미지 분류와 객체 검출 등 이미지 분석과 관련된 전반적인 분야에서 우수한 성능을 보이고 있으며, 최근에도 ‘ImageNet’ 이

나 'MSCOCO' 등 이미지 딥러닝 대회를 통해 우수한 성능의 이미지 딥러닝 모델들이 꾸준히 공개되고 있다(Alex et al, 2012; Christian et al, 2015; He et al, 2016).

이미지 캡셔닝은 이미지를 이해하고 그에 적합한 캡션을 생성하는 기술이며, 이미지 인덱싱 및 검색, 의학, 심리학, 교육, 그리고 소셜 미디어 등 다양한 분야에서 활용될 수 있다. 이미지 캡셔닝의 기본 동작 원리는 <Figure 4>를 통해 설명할 수 있다. <Figure 4>의 상단은 CNN과 LSTM을 활용한 학습 과정을, 하단은 추론 과정을 나타내고 있다. 학습 단계에서는 이미지와 캡션이 동시에 입력으로 사용되며, 디코더에서는 캡션의 각 단어들이 각 단계 LSTM 학습의 입력으로 사용된다. 한편 추론 단계에서는 캡션이 없는 이미지가 입력으로 사용되며, 디코더에서는 이전 LSTM의 출력이 다음 LSTM 학습의 입력으로 사용된다. 이때 각 LSTM의 출력을 조합한 것이 해당 이미지의 최종 결과물인 캡션이다.

다양한 측면에서 이미지 캡셔닝의 성능을 향상시키고자 하는 시도가 꾸준히 이루어지고 있으며, 특히 단어를 생성할 때마다 이미지의 특정 영역에 집중하는 어텐션 기법을 이미지 캡셔닝에 적용하면서 그 성능이 한 단계 도약하였다. <Figure 5(a)>는 어텐션 기반 이미지 캡셔닝(Xu et al, 2015)의 예로, 좌측 이미지로부터 밑줄 친 단어를 생성할 때 가장 집중했던 영역이 우측 이미지에서 흰 색으로 표시되어 있다. 또한 하나의 이미지에 포함된 다수의 객체 각각에 대해 캡션을 생성하는 'Dense Captioning'(Justin et al, 2016)에 대한 연구를 비롯하여, 다양한 관점에서 이미지 캡셔닝을 고도화하기 위한 연구(Hossain et al, 2019)가 활발하게 이루어지고 있다.

최근에는 이미지를 단순히 정확하게 기술하는 것을 넘어, 이미지에 내재된 정보를 더욱 세련되게 전달할 수 있는 고급 캡션을 생성하기 위한 연구가 이루어지고 있다. <Figure 5(b)>는 이미지에 대해 원하는 스타일의 캡션을 생성하는

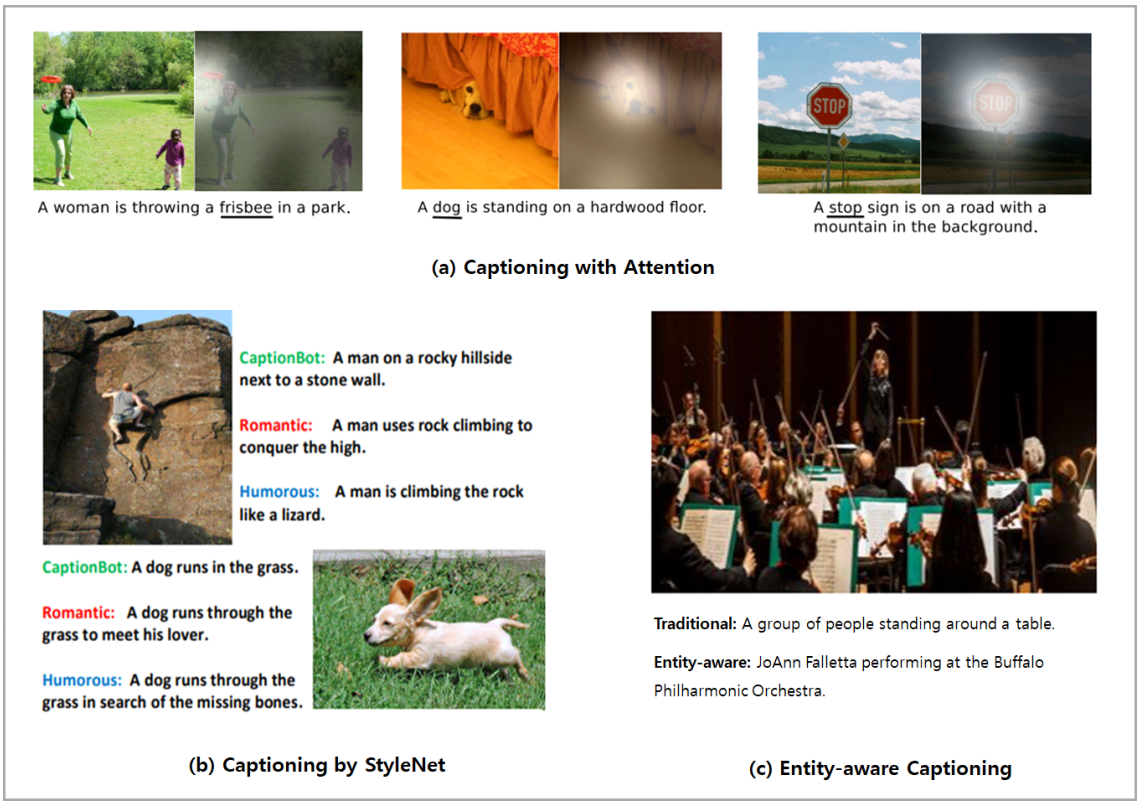


〈Figure 4〉 Training and Inference Process of Image Captioning

StyleNet(Gan et al, 2017)의 결과이다. <Figure 5(b)>에서 CaptionBot은 전통적인 캡셔닝 결과를, Romantic과 Humorous는 각기 다른 스타일을 적용한 캡셔닝 결과를 보이고 있다. 한편 이미지의 출처로부터 추가 정보를 수집하여 더욱 구체적인 캡션을 생성하는 연구(Ali et al, 2019)도 수행되었다. <Figure 5(c)>에서 Traditional은 일반적인 서술로 구성된 캡션을 나타낸다. 이와 달리 Entity-aware Captioning(Ali et al, 2019)은 이미지가 포함되어 있던 문서로부터 'a group of people'이 오케스트라라는 정보와 지휘자의 이름이 'JoAnn Falletta'라는 추가 정보를 획득한 후, 이를 활용하여 더욱 구체적인 캡션을 생성한다.

2.3 전이 학습

전통적으로 머신러닝 혹은 딥러닝 학습 알고리즘은 특정 과제나 문제에 국한하여 학습하도록 설계되었다. 이는 특정 도메인의 문제를 해결하기 위해 해당 도메인의 데이터가 필요하며 그 도메인에 맞는 모델이 설계되어야 함을 의미한다. 이와 달리 전이 학습은 사람이 여러 도메인과 과제를 넘나들면서 지식을 활용하는 것처럼, 특정 도메인에서 학습한 지식을 다른 도메인 혹은 과제에서 활용할 수 있도록 모델을 재사용하는 방법을 일컫는다. 전이 학습 방법은 데이터의 부족 문제를 해결해 줄 수 있을 뿐 아니라 모델



〈Figure 5〉 Recent Researches in the Literature of Image Captioning

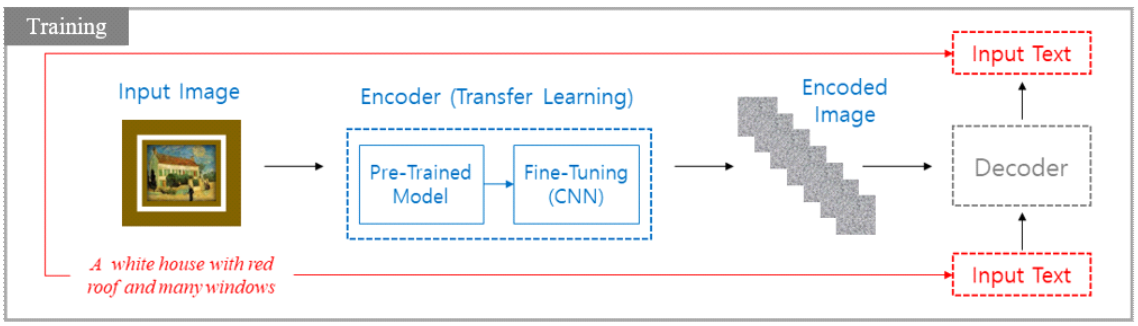
의 성능을 높이는 데 중요한 역할을 한다. 특히 학습에 대량의 데이터가 반드시 필요한 딥러닝의 한계를 상당 부분 해결할 수 있으며, 높은 재사용성, 성능의 고도화, 그리고 데이터 부족 문제의 해결 등의 장점으로 인해 최근 많은 주목을 받고 있다.

전이 학습은 딥러닝 분야에서 사전 학습 모델(Ian and Yoshua, 2016; Karl et al, 2016; Pan et al, 2010; Tan et al, 2018)이라는 형태로 활용이 되며, 사전에 대량의 데이터로부터 학습된 모델의 가중치를 새로운 도메인에 재사용 하는 방식으로 수행된다. <Figure 6>은 전이 학습을 활용한 이미지 캡셔닝 모델의 학습 과정을 나타내고 있다. 이미지의 특징을 추출하는 Encoder 부분에서 대량의 데이터를 통해 사전 학습된 모델이 사용되었으며, 사전 학습 모델을 통해 이미지의 일반적인 특징을 가중치로 추출하고, 추출된 가중치를 타겟 도메인에 맞게 미세하게 조정하는 과정을 거치게 된다. 사전 학습 모델을 통해 데이터가 부족한 도메인에서도 이미지의 일반적인 특성을 추출할 수 있고, 미세조정을 통해 도메인에 적합한 세부 특성들을 추출할 수 있다. 하지만 전이 학습을 지나치게 상이한 도메인에 적용하는 경우 오히려 성능이 떨어지는 한계가 발견되고 있고, 이런 한계를 해결하기 위한 연구들이 꾸준히 이어지고 있다(Qi et al, 2020).

2.4 데이터 분석을 활용한 미술치료

미술 치료는 그림을 매개로 인간의 심리를 진단하고 치료하는 비언어적 심리 치료 방법이다. 정형화되기 힘든 인간의 심리, 특히 무의식은 구체적인 언어보다 오히려 비 언어적인 형태를 통해 직접적으로 나타나는 것으로 알려져 있으며, 이 점에 착안한 방법이 미술 치료이다. 미술 치료는 자신의 감정이나 생각을 구체적인 언어로 표현하는데 어려움을 겪는 아이들의 심리 치료에 주로 활용된다. 가장 대표적인 미술 치료 방법으로는 HTP(House-Tree-Person) Test(Buck, 1948)가 있다. HTP Test는 집, 나무, 사람의 그림을 통해 개인의 의식적, 무의식적 심리 상태를 파악하는 검사로, 최근까지 기본적인 미술 치료로 널리 활용되고 있다. HTP Test를 통해 미술 치료사는 아이들의 그림에 나타난 상징적인 의미를 해석하고, 그림에 투영된 아이들의 심리 상태를 진단하여 심리 치료에 활용한다.

하지만 미술 치료는 그림을 해석하는 과정에서 치료사의 주관과 경험에 따라 해석에 차이가 발생할 수 있으며, 동일한 환경이더라도 진단과



〈Figure 6〉 Training of Image Captioning Using Transfer Learning

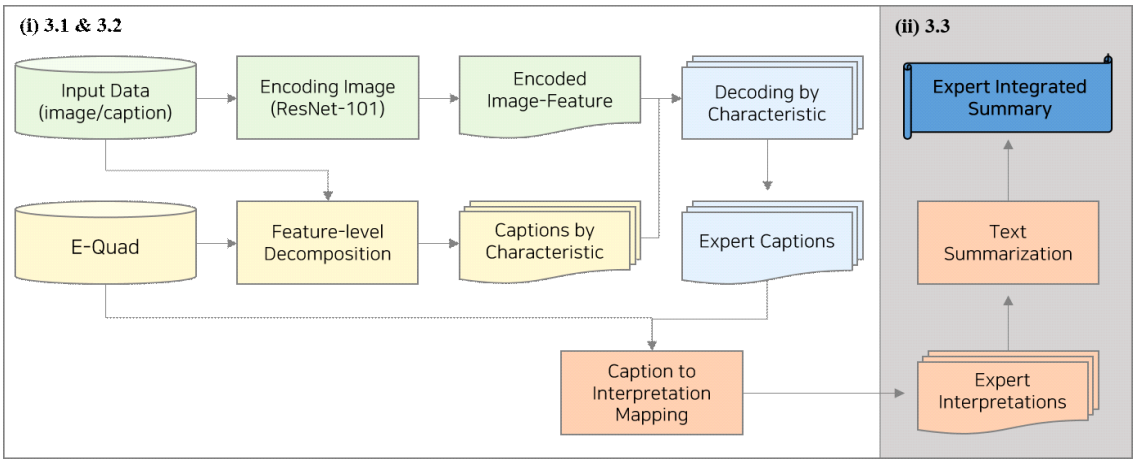
해석의 내용이 상이해질 수 있다는 한계를 가지고 있다. 이런 문제는 미술 치료뿐 아니라 심리학 연구 전체에서 재현성 위기(Kim et al, 2017)로 지적되고 있으며, 이는 미술 치료에 대한 일반인들의 신뢰도를 저하시키는 요인이 된다. 이를 해결하기 위해 미술 치료의 일부 과정을 인공지능으로 보완하여 객관성을 높이고자 하는 연구가 이루어진 바 있다. 이러한 맥락에서 본 연구에서는 미술 치료사의 그림 해석을 지원하기 위해, 입력 이미지에 대해 미술 치료 관점에서의 전문적 캡션을 생성하는 기법을 제안하고자 한다.

3. 제안 방법론

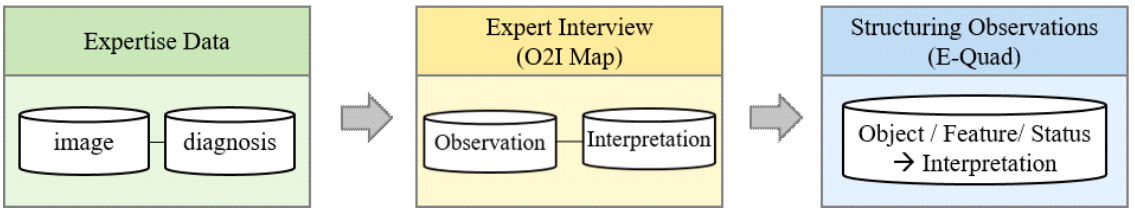
본 연구의 제안 방법론은 (1) 관찰/해석 지도 및 전문성 쿼드 생성, (2) 특성 독립 전이 학습 모델, (3) 전문적 통합 해석 생성의 세 단계로 구성된다. 아래 <Figure 7>은 제안 방법론의 전체 개요를 나타내고 있다.

3.1 관찰/해석 지도(O2I Map) 및 전문성 쿼드(E-Quad) 생성

첫 단계에서는 전문가와의 인터뷰를 통해 관찰/해석 지도(O2I Map: Observation to Interpretation Map)와 전문성 쿼드(E-Quad: Expertise Quadruple)를 생성한다. 구체적으로 이 과정은 미술 치료에 사용된 그림과 이에 대한 진단을 수집하는 전문 데이터 확보, 각 해석이 어떤 관찰을 근거로 이루어졌는지를 확인하는 전문가 인터뷰, 그리고 각 관찰을 객체(Object), 특성(Feature), 상태(Status)로 구조화하는 관찰 구조화의 과정으로 구성되며, 이 과정은 <Figure 8>에 나타나 있다. 예를 들어 어떤 그림에 대한 미술 치료사의 진단이 “창문을 많이 그린 것으로 보아, 아이가 외부와 접촉하고자 하는 강한 욕구를 갖고 있음을 알 수 있다”로 나타났다면, 이는 전문가 인터뷰를 통해 O2I Map에서 “창문의 개수가 많다”의 관찰과 “외부와 접촉하고자 하는 욕구가 강하다”의 해석으로 분리될 수 있다. 더 나아가 이러한 관찰이 ‘창문’이라는 객체, ‘개수’라는 특성, 그리고



〈Figure 7〉 Overview of the Proposed Model



〈Figure 8〉 O2I Map and E-Quad

〈Table 1〉 A Simple Example of E-Quad

Object	Feature	Status	Interpretation
Window	Count	Many	A desire to be in contact with the outside
		No	Closed tendency
Window	Shape	Oval	Progressive attitude
Window	Size	Small	Psychological distance or shyness

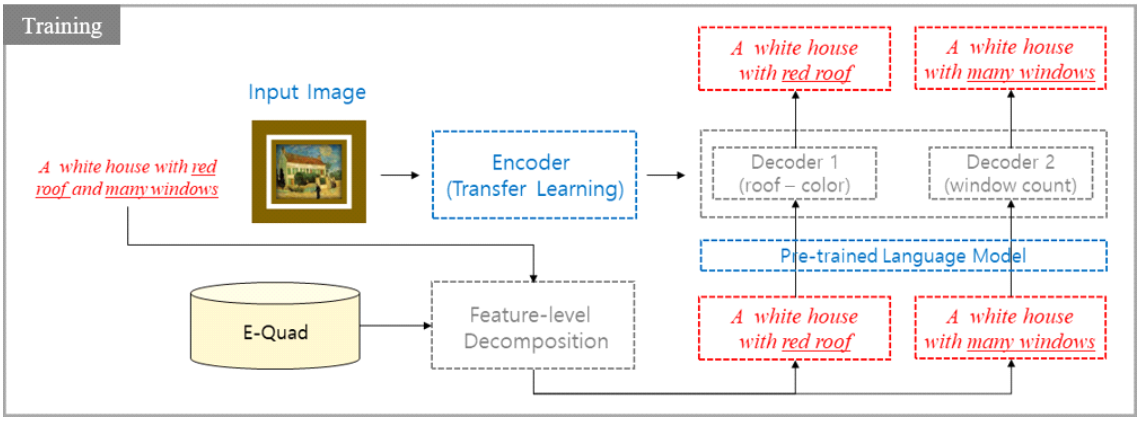
‘많다’라는 상태로 구조화 되어 해석과 함께 E-Quad를 구성한다(Table 1).

<Table 1>은 관찰 구조화의 이해를 돕기 위해 전문가의 자문을 받아 간략하게 작성한 E-Quad의 일부이며, 창문의 개수, 모양, 크기 등의 특성과 상태의 일부를 표현하고 있다. 이외에도 창문에 대한 추가 특성으로 위치, 색, 그리고 개폐 상태 등에 대한 내용들이 포함될 수 있으며, 창문 이외에도 문, 굴뚝, 지붕과 같은 객체들이 E-Quad에 포함될 수 있다.

3.2 특성 독립 전이 학습 모델

본 단계는 제안 방법론의 핵심 단계로, E-Quad를 참조하여 각 특성별로 별도의 캡션을 생성하고 미세 조정을 진행하는 과정을 나타낸다. <Figure 9>는 제안 방법론의 학습 과정을 도식화한 것으로, <Figure 6>의 전이 학습을 변형하여 본 연구에서 새롭게 제안하였다. 이해를 돕기 위해 본 그림에서는 하나의 문장만을 예로 들어 처리 과정을 소개한다. <Figure 9>에서 입력 텍스트는 E-Quad를 참조하여 각 특성별로, 즉 지붕(roof)에 대한 캡션과 창문(window)에 대한 캡션으로 재구성된다. 디코더는 각 특성을 처리하기 위한 독립 디코더들의 집합으로 구성되며, 각 특성에 대한 캡션들만 사용하여 미세 조정을 진행한다. 예를 들어 캡션 ‘A white house with red roof’는 ‘roof-color’ 특성을 처리하는 디코더의 입력으로 사용되며, ‘A white house with many windows’는 ‘window-count’ 특성을 처리하는 디코더의 입력으로 사용된다. 이와 같이 각 디코더는 하나의 특성에 대한 캡션만을 독립적으로 학습하므로, 제안 방법론은 앞에서 소개한 관찰간 간섭 현상을 효과적으로 방지할 수 있다.

한편 <Figure 9>의 인코더는 <Figure 6>과 같



〈Figure 9〉 Interference Preventing Transfer-learning

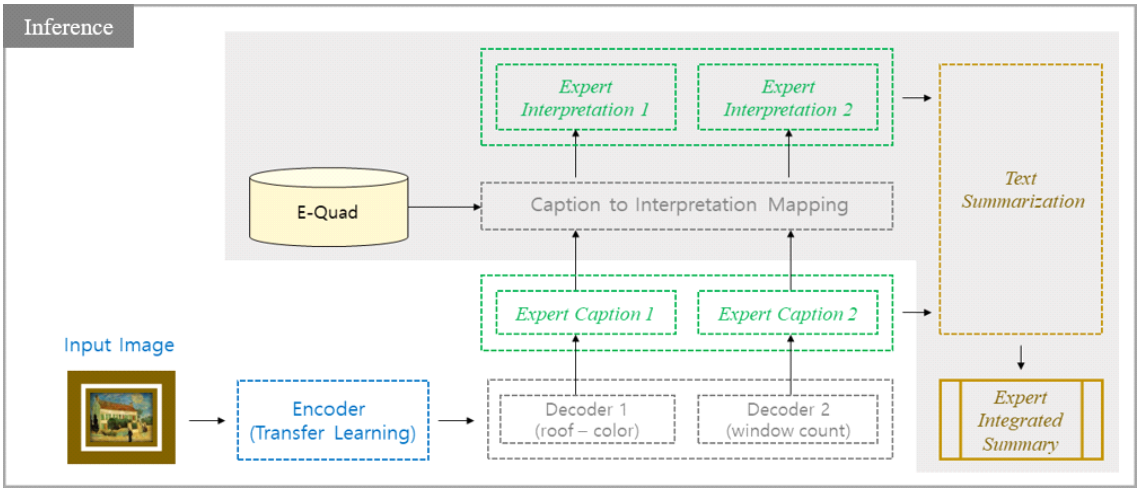
이 사전 학습 모델과 미세 조정의 두 가지 모듈로 구성되므로, 전문 데이터에 대한 미세 조정에 앞서 충분한 양의 일반 이미지/캡션 데이터에 대한 사전 학습이 이루어져야 한다. 사전 학습을 위한 대량의 데이터로는 Imagenet, MSCOCO 등의 이미지 셋이 널리 활용되고 있으며, 이미지 모델로는 Densenet(Huang et al, 2017), InceptionV3 (Christain et al, 2015), ResNet(He et al, 2016) 등이 많은 연구에서 사용되고 있다. 본 연구에서는 MSCOCO 데이터를 사용하고 Residual Block과 Skip-Connection의 특징을 가지며, 딥러닝의 깊은 레이어에도 정보의 손실이 적다는 강점을 갖는 ResNet-101 모델을 채택하여 실험을 진행한다.

<Figure 9>에서 전이 학습을 위한 인코더는 세부적으로 저수준 모듈과 고수준 모듈로 구성되어 있다. 저수준 모듈은 ResNet-101 모델을 통해 일반적인 이미지의 특성을 추출한다. 고수준 모듈은 추출된 저수준의 특성을 기반으로 전문 데이터에 대한 미세 조정을 수행하여 전문적인 특성을 추출하는 역할을 한다. 또한 텍스트 디코딩에는 어텐션을 접목한 LSTM 모델이 사용된다.

3.3 전문 해석 캡션 생성 모델

제안 방법론의 마지막 단계에서는 이전 단계에서 특성별로 생성된 캡션 및 E-Quad로부터 특성별 해석을 도출하고, 특성별 캡션과 해석을 통합하여 해당 이미지에 대한 전문적인 통합 해석을 생성한다. 이 과정은 <Figure 10>에 나타나 있으며, 그림에서 밝게 표시된 부분은 이전 단계에서 학습된 모델을 통해 각 특성별 캡션을 추론하는 과정을, 그리고 어둡게 표시된 부분은 추론된 캡션으로부터 해석을 도출하고 통합 해석을 생성하는 과정을 나타낸다.

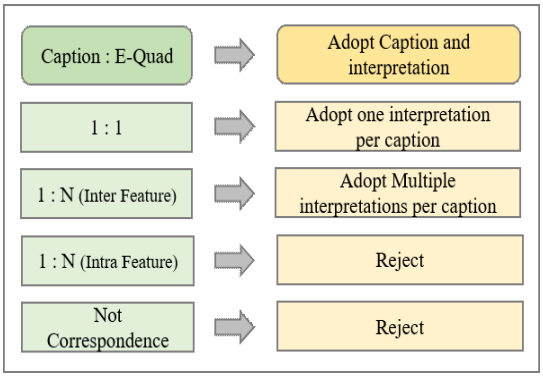
<Figure 10>에서 각 특성별로 추론된 캡션으로부터 해석을 생성하고 통합하는 과정은 크게 두 단계로 이루어진다. 우선 첫 단계는 각 캡션과 E-Quad의 매핑을 통해 각 캡션에 대한 미술치료 관점의 전문적 해석을 생성하는 단계이고, 두 번째는 생성된 캡션과 해석의 집합을 하나의 문서로 통합하는 단계이다. 두 번째 단계의 경우 (A) 캡션과 해석의 쌍으로 구성된 테이블 생성, (B) 캡션과 해석의 단순 통합, (C) 텍스트 요약



〈Figure 10〉 Generation and Integration of Interpretation

(Text Summarization) 등 다양한 방법으로 비교적 용이하게 구현할 수 있다. 한편 첫 단계인 캡션과 E-Quad의 매핑은 기본적으로 캡션에 포함된 유효 토큰(객체, 특성, 상태)의 탐색을 통해 이루어진다. 예를 들어 캡션 'A white house with many windows'는 ['A', 'white', 'house', 'with', 'many', 'windows']의 토큰들로 분할되며, 이들 중 E-Quad에 수록된 객체 'window'와 상태 'many'가 유효 토큰으로 식별된다. 또한 이 두 정보의 조합으로 해당 캡션의 특성이 'count'임을 판단할 수 있으며, 결과적으로 이 캡션은 E-Quad의 'window-count-many'에 대응되는 해석을 생성한다.

캡션과 E-Quad의 대응은 여러 형태로 이루어질 수 있으며 이는 <Figure 11>에 나타나 있다. 하나의 캡션이 하나의 E-Quad에 대응되는 경우는 대응되는 해석을 해당 캡션의 해석으로 채택하고, 어떤 E-Quad에도 대응되지 않는 캡션은 별도의 해석을 생성하지 않는다. 한편 둘 이상의 E-Quad에 동시에 대응되는 캡션이 존재하는 경우, 이러한 중복 대응이 동일 특성 내에서 발생하는지 특성 간 발생하는지 여부에 따라 처리가 달라진다. 즉, 특성 간 중복은 하나의 캡션이 여러 특성을 기술하는 경우를 의미하며, 이러한 캡션은 복수의 해석으로 대응된다. 한편 특성 내 중복은 동일한 특성에 대해 상이한 상태를 기술하는 것이므로, 해석이 불가능한 것으로 판단하여 별도의 해석을 생성하지 않는다.



〈Figure 11〉 Interpretation based on Mapping between Caption and E-Quad

4. 실험

4.1 실험 개요

본 실험은 전문 해석 캡션의 이전 단계로, 전문가가 주목하는 각 특성에 대한 '전문 관찰 캡션'을 분리하여 생성하는 과정을 다룬다. 먼저 소량의 전문 데이터로 제안 모델이 관찰 속성(객체, 특성, 상태)을 잘 분리하여 학습할 수 있도록 데이터의 특징(이미지 및 캡션의 복잡도)에 따른 캡션 품질을 평가하는 실험을 진행하며, 이를 활용하여 '전문 관찰 캡션'을 생성하는 실험을 진행한다. 그 후 일반적인 캡션을 학습한 모델과 전문적인 캡션을 학습한 모델이 동일한 이미지에 대해 추론한 캡션의 결과를 비교하는 실험을 진행한다.

구체적인 실험 모델은 PyTorch를 기반으로 구현했으며, 이미지 인코딩을 위한 사전 학습 모델은 ResNet-101을 이용했다. 또한 어텐션 모델의 적용은 Show attend and tell(Xu et al, 2015)의 모델을 참고했으며, Attention Map을 이용하여 제안 모델이 관찰 특성을 제대로 추출하여 학습하는지를 확인하였다. 일반 캡션 학습에는 MSCOCO의 2014년 데이터를 사용하였으며, 구체적으로 훈련용(Training) 데이터 약 8만 건, 검증용(Validation) 데이터 약 4만 건, 그리고 이미지 각각에 대해 5개씩 부여된 캡션을 실험에 사용하였다.

전문 캡션 학습은 미술 치료사의 전문성 이식을 주제로 수행했으며, 하나의 관찰 속성(window-count)에 초점을 맞춰 진행하였다. 학습 데이터의 특징에 따른 결과의 차이를 파악하기 위해 이미지/캡션 데이터를 여러 버전으로 나누어 생성하였으며, 이미지는 창문이 없는 집의 이미지를 기본으로 하여 임의의 형태의 창문 임의의 개수를 임의의 위치에 추가하는 방식으로 생성했다. 가장 단순한 특징을 가지는 version1에서부터 가장 복잡한 특징을 가지는 version4까지 복잡도를 변화시켜가며 다양한 학습 이미지를 생성하고, 각각에 대해 반복 실험을 수행하였다. <Figure 12>는 버전에 따른 이미지 데이터의 예시이며, 아래 <Table 2>는 버전별 데이터의 특징을 간략히 요약하고 있다.

또한 캡션은 객체명과 상태명의 단순 조합인 'Two words caption'과 객체명과 상태명을 각각 window와 no/one/many로 통일하고 그 외의 요소는 별도의 제약 없이 자유롭게 표현하는 'Sentence caption'으로 나누어 생성했다.

〈Figure 12〉 Example of Images in Each Experiment

〈Table 2〉 Characteristics of Each Experiment

Version	House Type	Background	Window Type	Window Size
1	Fixed	Fixed	Fixed	Fixed
2	Fixed	Fixed	Fixed	Varied
3	Fixed	Fixed	Varied	Varied
4	Varied	Varied	Fixed	Varied

4.2 이미지 복잡도에 따른 캡션 품질 비교

이미지의 복잡도에 따라 다르게 생성된 이미지/캡션 데이터로부터 각각 모델을 생성하고, 이를 통해 생성한 추론 결과를 비교하였다. <Figure 13>은 왼쪽 이미지로부터 추론할 수 있는 다양한 유형의 가상 결과들이며, 문장의 완성도와 객체 인식의 정확도에 따라 다양한 결과가 도출될 수 있음을 보이고 있다. (a)는 실제 이미지와 동일하게 창문을 하나로 묘사했다는 점에서 정확도는 높지만, ‘one’이 두 번 연속으로 사용되었다는 점에서 문장의 완성도는 낮다고 할 수 있다. 이와 유사하게 (b)는 정확도와 문장의 완성도가 모두 낮은 경우, (c)는 정확도는 낮지만 문장의 완성도가 높은 경우, 그리고 (d)는 정확도와 문장의 완성도 측면 모두에서 품질이 높은 캡션을 생성한 경우를 나타낸다.

실험 결과는 위에서 소개한 객체 특성 인식의 정확도와 문장의 완성도 외에 Attention Map을 활용하여 평가할 수 있다. Attention Map은 추론 과정에서 생성된 캡션의 각 단어가 이미지의 어떤 부분에 집중하였는지를 시각적으로 나타내며, 강조된 부분과 실제 단어의 일치 여부를 판단하여 모델의 성능을 직관적으로 평가할 수 있다. 아래 각 부절에서는 실험을 위해 새로 조합하여 생성한 63개 이미지에 대한 버전별 실험 결과를 소개한다.

4.2.1 생성 데이터: Version 1

Version 1의 실험 이미지는 창문과 집의 종류를 고정하고 창문의 위치와 개수만 임의로 변경

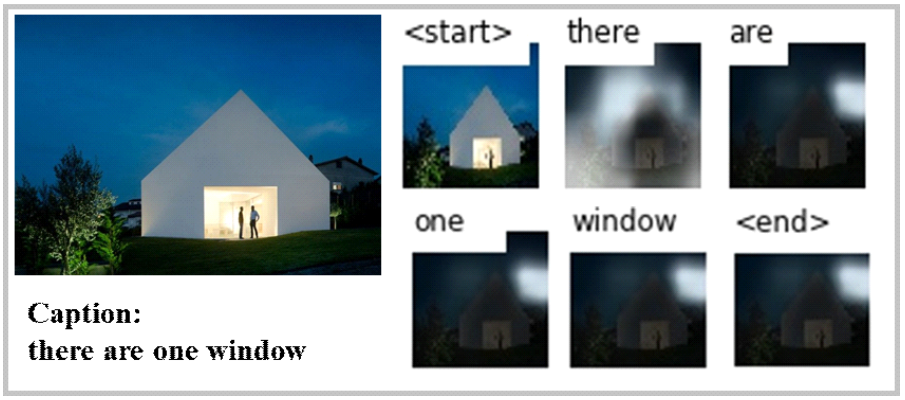
Expected Caption	Accuracy	Completeness
(a) there are one one window	–	–
(b) there is is many windows	–	–
(c) there are many windows	–	–
(d) there is one window	–	–

〈Figure 13〉 Imaginary Example of Expected Results

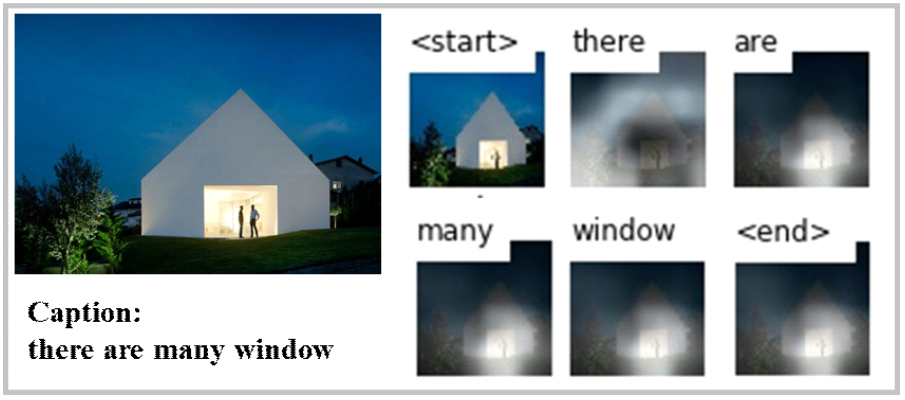
하여 생성하였으며, 총 51개의 학습 데이터와 13개의 검증 데이터를 사용했다. 생성된 모델로 추론을 진행한 결과, 63개의 이미지 중 count를 바르게 생성한 데이터가 33개로 정확도는 약 52.3%로 나타났다. <Figure 14>는 왼쪽 이미지에 대해 Version 1의 데이터로 생성한 모델을 사용한 추론 결과 중 하나로, 문장의 완성도가 낮고 Attention Map 또한 단어와 관련이 없는 엉뚱한 곳을 강조하고 있는 것을 확인할 수 있었다. 이러한 현상은 다른 이미지에 대한 추론 대부분에서 유사하게 나타났다.

4.2.2 생성 데이터: Version 2

Version 2는 Version 1과 마찬가지로 창문과 집의 종류는 고정하되 창문의 크기를 변화시킨 이미지를 사용하였으며, 마찬가지로 51개의 학습 데이터와 13개의 검증 데이터로 모델을 생성했다. 추론 결과 63개의 이미지 중 count를 바르게 생성한 데이터가 37개로 정확도는 약 58.3%로 나타났다. Version 1에 비해 다소 높은 객체 인식 정확도를 보였지만, 문장의 완성도와 Attention Map 일치도는 여전히 낮음을 확인하였다(Figure 15).



〈Figure 14〉 A Result of Caption Generation with Version 1



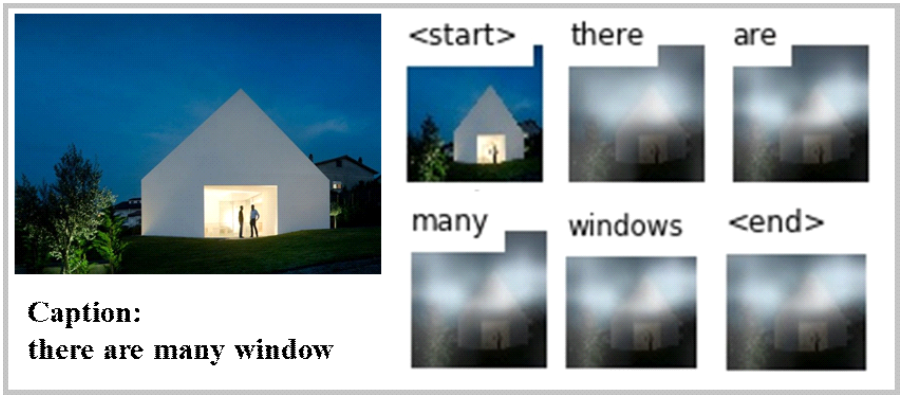
〈Figure 15〉 A Result of Caption Generation with Version 2

4.2.3 생성 데이터: Version 3

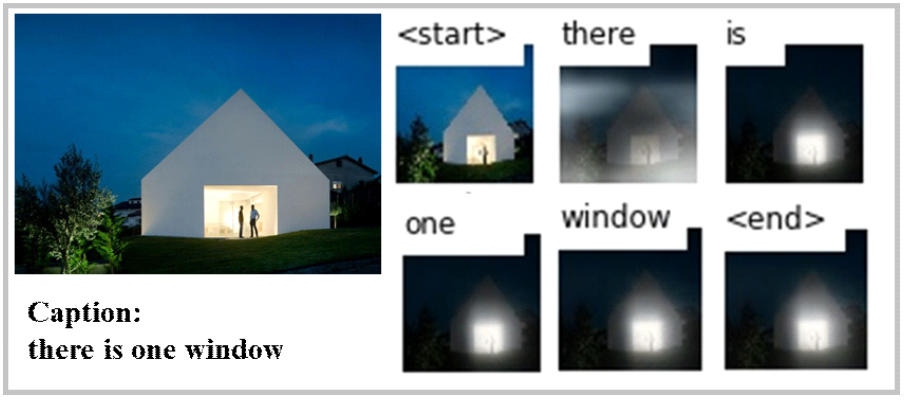
Version 3은 집의 종류는 하나로 통일하되 창문의 종류를 2가지로 변경하며 데이터를 생성하였으며, 총 67개의 학습 데이터와 19개의 검증 데이터로 모델을 생성하고 추론했다. 추론 결과 63개의 이미지 중 count를 바르게 생성한 데이터가 35개이고 정확도는 약 55.5%로 이전 실험들과 유사하게 나타났지만, 문장의 완성도는 다소 높게 나타났다(Figure 16).

4.2.4 생성 데이터: Version 4

Version 4는 창문의 종류와 크기는 고정하고 창문 이외에 집과 배경을 11가지로 다양하게 사용하였으며, 총 76개의 학습 데이터와 25개의 검증 데이터를 활용해 모델을 생성했다. 생성된 모델로 63개의 이미지를 추론한 결과, count를 바르게 생성한 데이터가 52개, 정확도가 약 82.5%로 이전 실험들에 비해 상대적으로 높은 정확도를 나타냄을 확인할 수 있었다. 또한 문장의 완성도도 비교적 높게 나타났으며, Attention Map의 경우도 각 단어에 해당되는 이미지 영역을 제대로 강조하고 있음을 확인하였다(Figure 17).



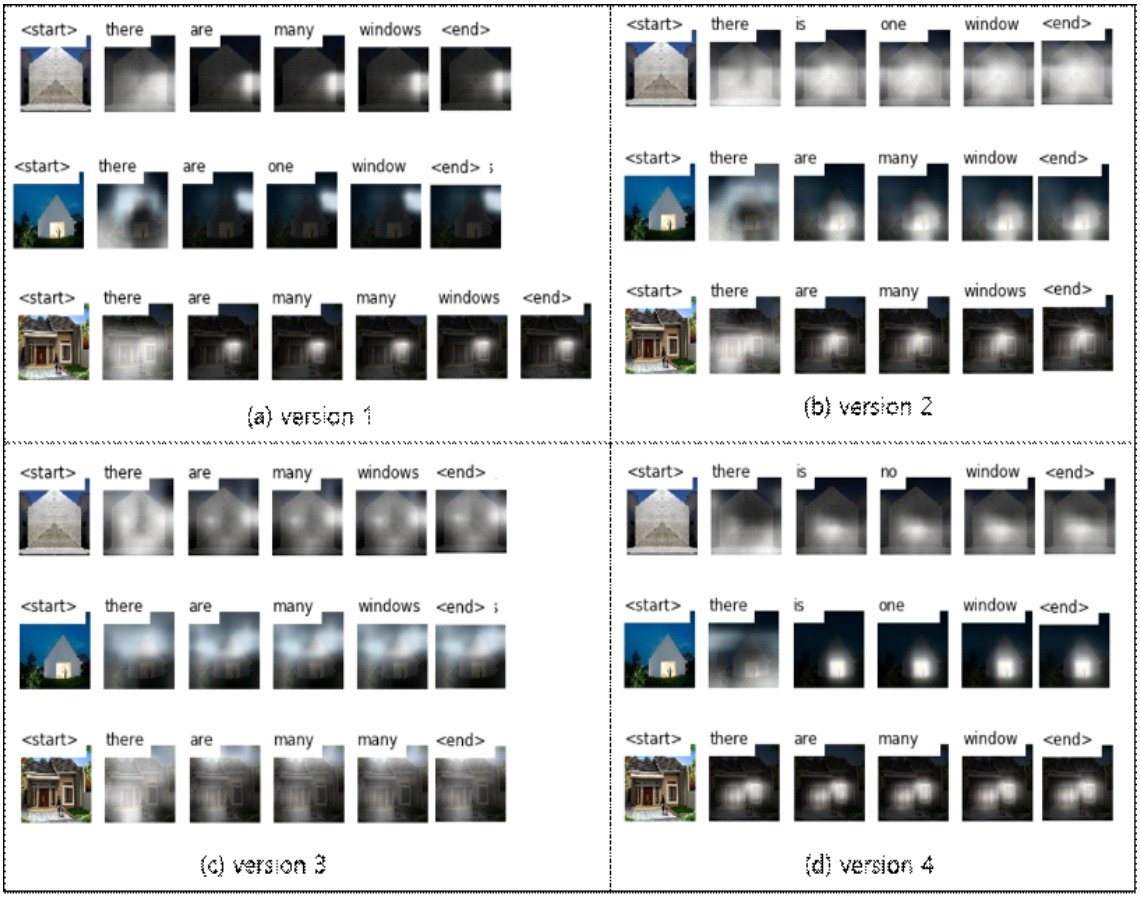
〈Figure 16〉 A Result of Caption Generation with Version 3



〈Figure 17〉 A Result of Caption Generation with Version 4

4.2.5 결과 해석

소량의 전문 캡션 데이터를 이용해 전문 캡션 모델을 생성하는 경우, 데이터의 편향이 학습에 많은 영향을 미치게 된다. 따라서 데이터의 편향을 줄이고 학습에 악영향을 주는 요소를 제거하기 위해 이미지 복잡도에 따른 캡션의 품질 비교 실험을 진행했다. <Figure 18>은 버전별 모델을 적용한 실제 추론 결과의 일부를 보이며, 이를 통해 학습 이미지의 복잡도에 따른 캡션의 품질을 비교할 수 있다. Version 1~3의 데이터를 이용한 실험 결과, 데이터가 지나치게 단순하여 모델이 데이터의 특성을 충분히 학습하지 못한 것으로 파악된다. 추론으로 생성한 캡션의 약 절반이 창문의 수를 정확하게 파악하지 못했고, Attention Map에서도 각 단어가 이미지의 관련된 부분에 제대로 집중하고 있지 못한 것을 확인할 수 있었다. 이는 곧 모델이 객체의 특성을 제대로 학습하지 못함을 의미한다. 반면 Version 4의 데이터를 이용한 실험은 객체 인식의 정확도와 문장의 완성도, 그리고 Attention Map의 일치도 측면 모



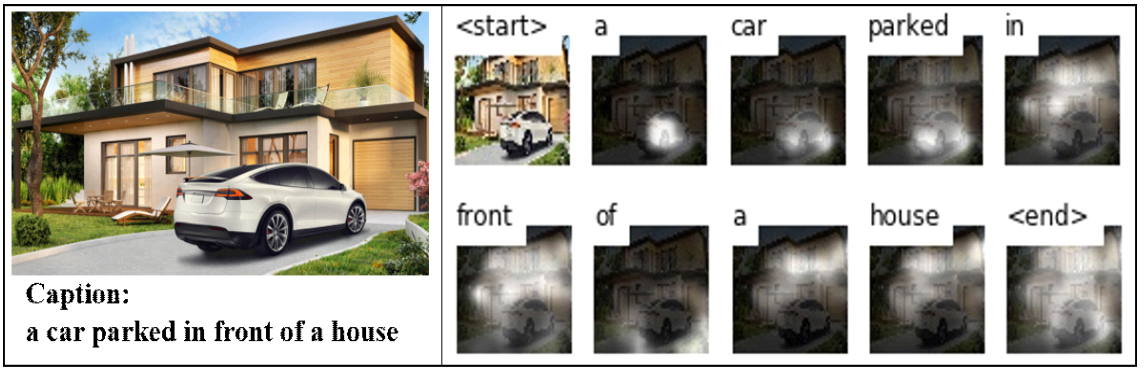
〈Figure 18〉 Some Results of Caption Generation

두에서 만족할 만한 결과를 도출하였다. Version 4는 관심 속성인 창문은 고정시키고 오히려 그 외의 요소를 다양하게 학습한 것으로, 본 실험을 통해 전문성 이식에 사용되는 데이터는 관심 속성과 비 관심 속성을 구분해서 활용할 필요가 있음을 확인하였다.

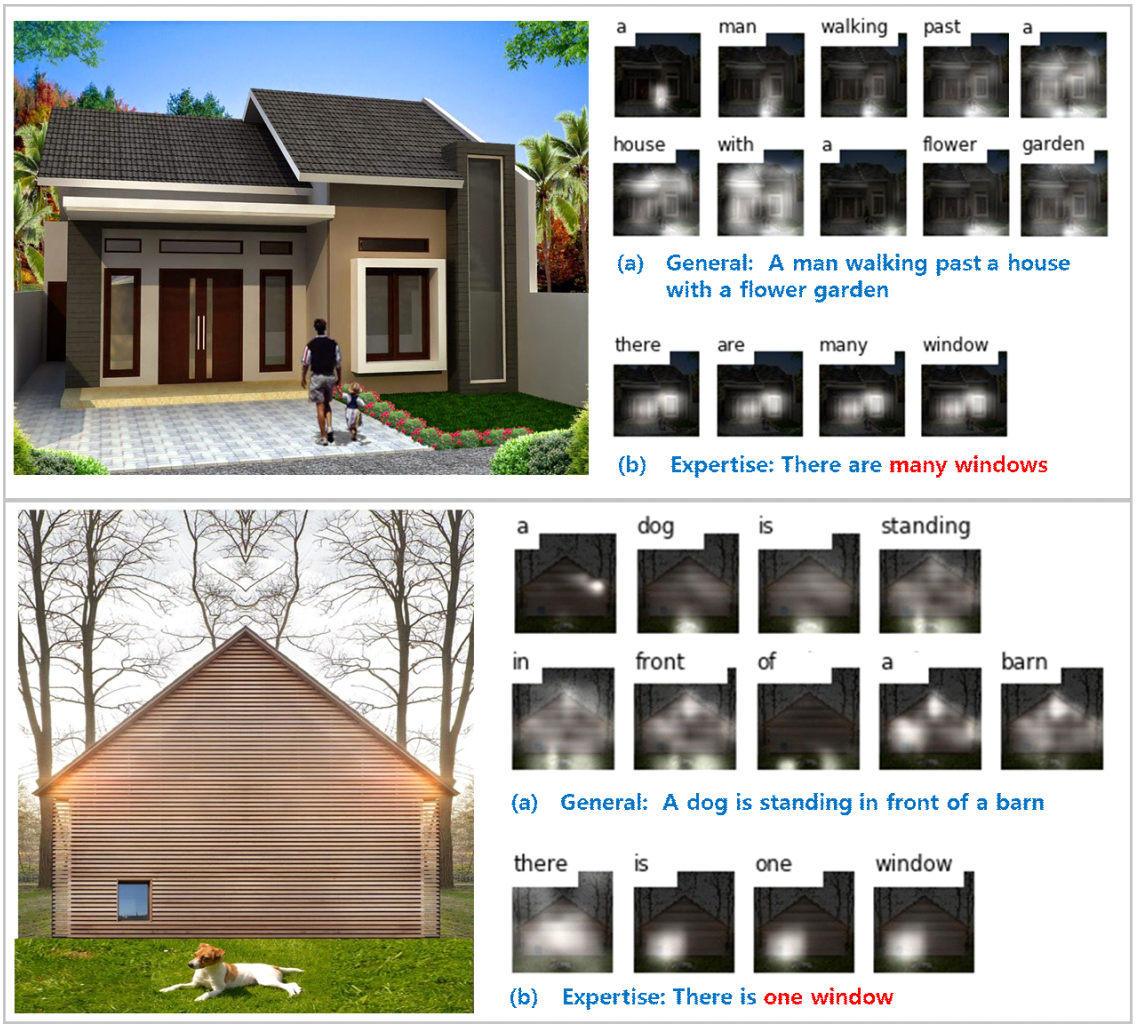
4.3 일반 캡션과 전문 캡션

본 절에서는 일반 캡션의 생성 결과와 전문 캡션의 생성 결과를 비교함으로써 제안 방법론의 우수성을 평가한다. <Figure 19>는 주어진 이미지에 대해 일반적인 관점을 학습한 사전 학습 모델을 사용하여 이미지를 추론한 결과를 보이고 있다. 주어진 이미지와 관련된 완성도 높은 문장을 생성하였지만, 생성된 캡션은 미술 치료 분야의 HTP Test와 관련된 요소에 대한 내용은 중요하게 다루고 있지 않다. 본 절의 이후 부분에서는 주어진 이미지에 대해 일반적인 관점에서 생성한 캡션과 미술 치료 전문가의 관점, 구체적으로는 창문의 개수에만 집중하여 생성한 캡션을 비교한다.

일반 캡션과 전문 캡션의 비교는 <Figure 20>에서 확인할 수 있다. 그림에서 (a) General은 사전 학습 모델을 그대로 적용하여 생성한 일반적인 관점의 캡션을 나타내며, (b) Expertise는 제안 방법론을 활용하여 창문의 개수에만 집중하여 생성한 전문 캡션을 나타낸다. 구체적으로 전문 캡션은 사전 학습 모델을 사용하여 이미지의 일반적인 특성을 학습한 뒤, 4.2절의 Version 4 유형의 전문 데이터에 대한 미세 조정을 통해 도출한 모델로부터 추론하였다. <Figure 20>에서 일반적 캡션을 학습한 모델은 이미지에 포함된 객체 식별을 중심으로 비교적 자연스러운 표현의 캡션을 생성하였으나, 전문적 해석에 중요하게 사용되는 객체와 특성을 누락하거나 전문적 해석과 무관한 내용을 포함하고 있는 것으로 나타났다. 한편 전문적 캡션을 학습한 모델은 소량의 이미지에 대해 미세 조정을 수행했음에도 불구하고, 전문적 해석에 중요하게 사용되는 객체와 특성 위주의 캡션을 완성도 높게 생성한 것을 확인할 수 있다.



〈Figure 19〉 General Caption inference



〈Figure 20〉 Preliminary Results of Expertized Image Captioning (Figure 3, Revisited)

5. 결론

본 연구에서는 주어진 이미지에 대해 관련 캡션을 자동으로 생성하는 기술인 이미지 캡셔닝을 더욱 고도화하기 위한 방안을 제시하였다. 이미지 캡셔닝의 성능을 고도화하기 위한 최근의 많은 노력에도 불구하고, 이미지를 일반인이 아닌 분야별 전문가의 시각에서 해석하기 위한 연구는 찾아보기 어렵다. 이에 본 연구에서는 전문가의 전문성을 활용하여 이미지에 대해 해당 분야에 특화된 캡션을 생성하기 위한 방안을 제안하였다. 또한 데이터의 편향을 줄이고 학습에 악

영향을 주는 요소를 제거하기 위해 이미지 복잡도에 따른 캡션의 품질 비교 실험을 진행하였으며, 그 결과를 토대로 전문 캡션 생성 모델을 구축하고 캡션을 추론한 결과를 제시하였다.

이미지 캡셔닝은 이미지 처리와 텍스트 처리 양 분야에 대한 깊은 이해를 필요로 하므로, 다른 분야에 비해 진입 장벽이 높아 큰 활용 가치에도 불구하고 충분한 연구가 이루어지지 못했다. 이러한 측면에서 첫째, 본 연구는 딥러닝 기술의 양 축인 이미지 딥러닝과 텍스트 딥러닝을 동시에 다룬다는 점에서 도전성을 갖는다. 딥러닝 관련 최근 대부분의 연구들은 다양한 딥러닝 알고리즘의 성능을 비교하거나, 최근 알고리즘을 적용하여 분류 정확도를 향상시키는 것을 목적으로 하고 있다. 하지만 딥러닝 방법론 자체의 변형을 통해 성능을 개선하거나 새로운 활용 분야를 찾기 위한 연구는 상대적으로 매우 드물다. 이러한 측면에서 둘째, 본 연구는 전문 이미지 해석이라는 새로운 연구 목표를 제안하고, 이를 달성하기 위해 최근 관심이 집중되고 있는 전이학습의 새로운 활용 방안을 제안했다는 점에서 그 기여를 인정받을 수 있을 것이다. 또한 전이학습은 한 분야의 학습 결과를 다른 분야의 추론에 활용하기 위한 기법으로, 주로 딥러닝 학습에 필요한 데이터의 부족 현상을 극복하고 학습에 소요되는 시간을 줄이기 위해 사용되고 있다. 하지만 본 연구에서는 특정 분야의 전문성을 구조화하고, 이를 통해 일반적 사전 학습의 결과에 해당 분야의 전문성을 이식하기 위한 방안으로 전이 학습을 사용하는 방안을 제시했다. 향후 제안 방법론을 다양한 분야의 전문성 이식에 적용하여 전문 데이터 부족 문제를 해결하고 성능을 고도화하기 위한 연구가 활발하게 이루어질 것으로 기대한다.

본 연구에서는 단 하나의 도메인, 그 중에서도 관찰 속성 중 일부분에 대한 캡션 생성 실험을 수행하였으며, 이는 본 연구의 한계로 지적될 수 있다. 추후 연구에서는 다양한 도메인의 다양한 객체 및 관찰 속성에 대한 캡션 생성 실험을 수행할 필요가 있다. 또한 더욱 정교하고 견고한 학습을 위해 각 전문 분야의 관찰 속성 미세 조정을 위한 이미지/캡션 데이터를 충분히 확보할 수 있는 방안에 대한 고찰이 이루어져야 한다.

참고문헌(References)

Alex, K., S. Ilya, and E. H. Geoffrey, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, Vol. 25, (2012), 1097-1105.

Ali, F. B., G. Lluis, R. Marcal, and D. Karatzas, "Good News, Everyone! Context Driven Entity-Aware Captioning for News Images," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2019), 12466-12475.

Ashnish, V., S. Noam, P. Niki, U. Jakob, J. Llion, N. G. Aidan, K. Lukasz, and P. Illia, "Attention is All You Need,", arXiv:1706. 03762, (2017).

Buck J.N., "The H-T-P test," Journal of Clinical Psychology, Vol 4, (1948), 151-159.

Caigny, A. D., C. Krsitof, W. D. B. Koen, and L. Stefan, "Incorporating Textual Information in Customer Churn Prediction Models Based on a Convolutional Neural Network," International Journal of Forecasting, (2019), 1-16.

Chen, L., T. Zhang, and Y. Chen, "Customer Purchase

Intent Prediction Under Online Multi-Channel Promotion: A Feature-Combined Deep Learning Framework," IEEE Access, Vol. 7, (2019), 112963-112976.

Christain, S., W. Liu, Y. Jia, S. Pierre, R. Scott, A. Dragomir, E. Dumitru, V. Vincent, and R. Andrew, "Going Deeper with Convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2015), 1-9.

Devlin, J., MW. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805, (2018).

Feng, M., T. Shaonan, C. Lee, and M. Ling, "Deep Learning Models for Bankruptcy Prediction Using Textual Disclosures," European Journal of Operational Research, Vol. 274, No. 2, (2019), 743-758.

Forrest, N. I., S. Han, W. M. Matthew, A. Khalid, J. D. William, and K. Kurt, "SqueezeNet: AlexNet-level Accuracy with 50x Fewer Parameters and <0.5MB Model Size," arXiv:1602.07360, (2016).

Gan, C., Z. Gan, X. He, J. Gao, and D. Li, "StyleNet: Generating Attractive Visual Captions with Styles," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2017), 3137-3146.

He, K., X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2016), 770-778.

Hossain, M. D. Z., S. Ferdous, F. S. Mohd, and L. Hamid, "A Comprehensive Survey of Deep Learning for Image Captioning," ACM Computing Surveys, Vol. 51, No. 6, (2019), 1-36.

Huang, G., Z. Liu, V. D. M. Laurens, and Q.W. Kilian, "Densely Connected Convolutional Networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2017), 4700-4708.

Ian, G., B. Yoshua., and C. Aaron, Deep Learning, MIT Press, United Strates, 2016.

Jeffrey, P., S. Richard., and D. M. Christopher, "Glove: Global Vectors for Word Representation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, (2014), 1532-1543.

Justin, J., K. Andrej, and F. Li., "Densecap: Fully Convolutional Localization Networks for Dense Captioning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2016), 4565-4574.

Karl, W., M. K. Taghi, and D. Wang, "A Survey of Transfer Learning," Journal of Big Data, Vol. 3, (2016) 1-40.

Kim, B. N., J. W. Choi, H. S. Ko, “Replication crisis in psychology: A review of its causes and solutions,” Korean Journal of Psychology: general, Vol. 36. No. 3, (2017), 359-396.

Lecun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," Neural Computation, Vol. 1, No. 4, (1989), 541-551.

Liu, Y. and L. Wu, "Geological Disaster Recognition on Optical Remote Sensing Images Using Deep Learning," Procedia Computer Science, Vol. 91, (2016), 566-575.

Marc, T., G. Albert, and P. C. Kenneth, "Transfer

Learning from Language Models to Image Caption Generators: Better Models may not Transfer Better," arXiv:1901.01216, (2019).

Micheal, I. J., "Attractor Dynamics and Parallelism in a Connectionist Sequential Machine," Artificial Neural Networks: Concept Learning, (1990), 112-127.

Pan, S. J. and Q. Yang, "A Survey on Transfer Learning," IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 10, (2010), 1345-1359.

Pang, G., X. Wang, F. Hao, J. Xie, X. Wang, Y. Lin, and X. Qin, "ACNN-FM: A Novel Recommender with Attention-based Convolutional Neural Network and Factorization Machines," Knowledge-Based Systems, Vol. 181, (2019), 1-13.

Peters, M. E., N. Mark, I. Mohi, G. Matt, C. Christopher, K. Lee, and Z. Luke, "Deep Contextualized Word Representations," arXiv: 1802.05365, (2018).

Piotr, B., G. Eduard, J. Armand, and M. Tomas, "Enriching Word Vectors with Subword Information," arXiv:1607.04606, (2016)

Qi D., L. S., J. Song, E. Cui, T. Bharti, A. Sacheti, “ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data,” arXive:2001.07966, (2020).

Ren, S., K. He, G. Ross, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Advances in Neural Information Processing Systems, Vol. 28, (2015), 91-99.

Ryan, K., S. Ruslan, and Z. Richard, "Multimodal Neural Language Models," in Proceedings of the International Conference on Machine Learning, Vol. 32, (2014), 592-603.

Sanjiban, S. R., M. Abhinav, G. Rishab, S. O. Mohammad, and P. V. Krishna, "A Deep Learning Based Artificial Neural Network Approach for Intrusion Detection," in Proceedings of the International Conference Mathematics and Computing, (2017), 44-53.

Hochreiter, S. and S. Jürgen, "Long Short-Term Memory," Neural Computation, Vol. 9, No. 8, (1997), 1735-1780.

Tan, C., F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A Survey on Deep Transfer Learning," arXiv:1808.01974, (2018).

Tomas, M., K. Chen, C. Greg, and D. Jeffrey, "Efficient Estimation of Word Representations in Vector Space," arXiv:1301.3781, (2013).

Tomas, M., S. Ilya, K. Chen, C. Greg, and D. Jeffrey, "Distributed Representations of Words and Phrases and their Compositionality," Advances in Neural Information Processing Systems, Vol. 26, (2013), 3111-3119.

Xu, K., J. Ba, K. Ryan, K. Cho, C. Aaron, S. Ruslan, S. Z. Richard, and B. Yoshua, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in Proceedings of the International Conference on Machine Learning, Vol. 32, (2015), 2048-2057.

Yang, Y., L. Zheng, J. Zhang, Q. Cui, Z. Li, and P. S. Yu, "TI-CNN: Convolutional Neural Networks for Fake News Detection," arXiv: 1806.00749, (2018).

Yang, Z., Z. Dai, Y. Yang, C. Jaime, R. S. Russ, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," arXiv:1906.08237, (2019).

Abstract

Deep Learning-based Professional Image Interpretation Using Expertise Transplant

Taejin Kim* · Namgyu Kim**

Recently, as deep learning has attracted attention, the use of deep learning is being considered as a method for solving problems in various fields. In particular, deep learning is known to have excellent performance when applied to applying unstructured data such as text, sound and images, and many studies have proven its effectiveness. Owing to the remarkable development of text and image deep learning technology, interests in image captioning technology and its application is rapidly increasing. Image captioning is a technique that automatically generates relevant captions for a given image by handling both image comprehension and text generation simultaneously. In spite of the high entry barrier of image captioning that analysts should be able to process both image and text data, image captioning has established itself as one of the key fields in the A.I. research owing to its various applicability. In addition, many researches have been conducted to improve the performance of image captioning in various aspects. Recent researches attempt to create advanced captions that can not only describe an image accurately, but also convey the information contained in the image more sophisticatedly.

Despite many recent efforts to improve the performance of image captioning, it is difficult to find any researches to interpret images from the perspective of domain experts in each field not from the perspective of the general public. Even for the same image, the part of interests may differ according to the professional field of the person who has encountered the image. Moreover, the way of interpreting and expressing the image also differs according to the level of expertise. The public tends to recognize the image from a holistic and general perspective, that is, from the perspective of identifying the image's constituent objects and their relationships. On the contrary, the domain experts tend to recognize the image by focusing on some specific elements necessary to interpret the given image based on their expertise. It implies that meaningful parts of an image are mutually different depending on viewers' perspective even

* Graduate School of Business IT, Kookmin University
** Corresponding Author: Namgyu Kim
School of Management Information Systems, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 136-702, Korea
Tel: +82-2-910-5425, Fax: +82-2-910-4017, E-mail: ngkim@kookmin.ac.kr

for the same image. So, image captioning needs to implement this phenomenon.

Therefore, in this study, we propose a method to generate captions specialized in each domain for the image by utilizing the expertise of experts in the corresponding domain. Specifically, after performing pre-training on a large amount of general data, the expertise in the field is transplanted through transfer-learning with a small amount of expertise data. However, simple adaption of transfer learning using expertise data may invoke another type of problems. Simultaneous learning with captions of various characteristics may invoke so-called 'inter-observation interference' problem, which make it difficult to perform pure learning of each characteristic point of view. For learning with vast amount of data, most of this interference is self-purified and has little impact on learning results. On the contrary, in the case of fine-tuning where learning is performed on a small amount of data, the impact of such interference on learning can be relatively large. To solve this problem, therefore, we propose a novel 'Character-Independent Transfer-learning' that performs transfer learning independently for each character.

In order to confirm the feasibility of the proposed methodology, we performed experiments utilizing the results of pre-training on MSCOCO dataset which is comprised of 120,000 images and about 600,000 general captions. Additionally, according to the advice of an art therapist, about 300 pairs of 'image / expertise captions' were created, and the data was used for the experiments of expertise transplantation. As a result of the experiment, it was confirmed that the caption generated according to the proposed methodology generates captions from the perspective of implanted expertise whereas the caption generated through learning on general data contains a number of contents irrelevant to expertise interpretation.

In this paper, we propose a novel approach of specialized image interpretation. To achieve this goal, we present a method to use transfer learning and generate captions specialized in the specific domain. In the future, by applying the proposed methodology to expertise transplant in various fields, we expected that many researches will be actively conducted to solve the problem of lack of expertise data and to improve performance of image captioning.

Key Words : Deep Learning, Expertise Transplant, Transfer-Learning, Image Captioning, Artificial Intelligence

Received : May 12, 2020 Revised : June 4, 2020 Accepted : June 20, 2020
Publication Type : Regular Paper Corresponding Author : Namgyu Kim

저 자 소 개

김 태 진
현재 국민대학교 비즈니스IT전문대학원 석사과정에 재학 중이며, 국민대학교 경영정보학부에서 학사 학위를 취득하였다. 한국지능정보시스템학회 학술대회 우수 논문상, NTIS 정보활용경진대회 과학기술정보통신부 장관상, 캡스톤디자인 경진대회 한국연구재단 이사장상 등을 수상하였다. 주요 관심분야는 텍스트 마이닝, 딥러닝, 이미지 캡셔닝 등이다.

김 남 규
현재 국민대학교 비즈니스IT전문대학원장 및 경영정보학부 교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사학위를 취득하였다. 한국지능정보시스템학회 부회장, 한국정보기술응용학회 부회장, 한국경영학회 상임이사, 한국경영정보학회 이사, 한국인터넷정보학회 이사를 역임하였다. 주요 관심분야는 텍스트 마이닝, 딥러닝, 데이터 모델링 등이다.