|  |
| --- |
| 딥러닝을 이용한 정지 영상 해설 생성  2010104038 김 진 우  2014104145 정 소 현  **요약**  4차 산업 혁명의 핵심 기술 중 하나인 딥 러닝을 활용해서 프로젝트를 진행하고자 한다. 딥 러닝이란 머신러닝의 일종으로, 기계가 사람처럼 스스로 데이터를 분석 및 학습하고, 특징을 추출하는 것을 뜻한다. ‘인공 신경망’이라고 하는 기계가 스스로 데이터를 분석하고 답을 내는 방식인, 인간의 뇌 구조를 바탕으로 한 알고리즘을 기반으로 하고 있다. 우리가 구현하고자 하는 소프트웨어는 딥 러닝으로 이미지를 어느 정도 해석해서 단어를 뽑아내고 그 단어들을 자연어 처리를 통한 문장으로 만들어 주는 소프트웨어를 설계하고 구현해보려고 한다. 딥 러닝이 많이 쓰이는 분야인 Image Processing( 영상 처리 )과 Natural Language Processing( 자연어 처리 ) 두 기술을 융합하고 Web을 통해 사용자에게 서비스를 해보는 것을 목표로 하고 있다. |

1. **서론**
   1. **. 연구배경**

최근의 IT, ICT 기술들은 대부분이 개인화, 자동화에 초점이 맞추어져 있다. 이도 그럴 것이 많은 것들이 이미 발명되어 있기 때문에 이를 이용해서 어떻게 하면 사용자에게 좀 더 편리한 서비스를 제공할 수 있을 것인가에 집중하게 되는 것은 당연하다. 빅데이터 시대로 접어들면서 미국의 하버드 비즈니스 리뷰는 21세기의 가장 섹시한 직업으로 Data Scientist( 데이터 과학자 )를 꼽을 만큼 데이터를 활용하고 학습시켜서 무언가를 창출하는 일은 지금 그리고 앞으로의 IT시대에 매우 유망한 일이다. 이런 매력적인 상황들에 끌려서 우리는 데이터를 다루는 기술 중 하나인 머신 러닝, 그리고 머신 러닝의 기술 중 하나인 딥 러닝 기술에 초점을 맞추어서 프로젝트를 진행해보려 한다.

처음에 프로젝트에 대해서 조사를 하던 중 시각장애인에게 영화를 설명할 수 있도록 스크립트를 만들어주는 소프트웨어에 대한 연구를 보고 이런 방식으로도 딥 러닝 기술이 쓰일 수 있겠구나 라는 생각을 하였다. 특정 이미지 영상에 대한 설명을 자동화하게 되면 미술관, 이미지 검색 등 많은 분야에서 기존의 방법보다 편리하게 쓰일 수 있다. 그래서 우리는 정지된 이미지를 기준으로 이미지 안의 물체, 사람, 사람의 감정 등을 인식해서 인식한 내용을 단어로 뽑아주고 뽑은 단어를 자연어 처리를 통해 문장을 만들어서 이미지에 대한 사람의 이해를 돕는 스크립트를 생성하는 소프트웨어를 구현하고자 한다.

* 1. **. 연구목표**

가장 큰 목표는 자동화에 있다. 기존에는 이미지에 대한 설명을 작성하는 사람의 주관에 의해 사람의 손으로 작성되어졌다면 이 소프트웨어는 이미지에 나와있는 Object들을 적당히 파악하고 그에 맞는 학습 결과로 이미지를 자동으로 설명해주는 차이가 있다. 이로 인해 이미지에 대한 객관적인 정보를 뽑을 수도 있고 사람의 인력을 거치지 않은 자동화 시스템을 만들 수도 있고, 뽑은 데이터들을 또 다른 작업의 데이터로 사용할 수도 있다.

입력된 이미지 데이터를 비지도 학습을 통해 학습한다. 비지도 학습이란 이미지 분석 등에 쓰이는 학습 방법으로써 필요한 사전 정보가 없고 입력에 대한 목표 결과 값이 주어지지 않은 경우, 수집된 데이터간의 유사성을 찾아 입력 데이터를 파악하는 학습 방법을 말한다. 이 학습방법을 통해 이미지에서 얻어진 단어들을 자연어 처리를 통해 문장으로 토큰화 시킨다. 이렇게 생성된 문장을 보고 사용자의 이미지에 대한 이해를 더욱 더 증진시키는 것을 목표로 한다.

1. **관련 연구**
   1. **. 물체 검출 (Object Detection)**

물체 검출은 주어진 사진이나 동영상에서 관심있는 물체들 (예를 들면 인물, 자동차, 동물 등)이 어디에 위치하는지 찾는 문제이다. 즉 이미지 상에서 물체가 존재하는 영역을 경계 상자 (Bounding Box)나 픽셀 마스크로 지정하는 것이 목표이다. 물체 검출은 크게 두가지 작은 문제를 풀면서 진행되는데, 하나는 사진 상에서 물체가 있을만한 영역들을 빠르고 정확하게 열거하는 부분이고, 다른 하나는 관심있는 물체를 잘 표현하는 Feature Map을 변별력 있게 학습하여 주어진 영역에 대상 물체가 있는지 판단하는 것이다. 물체 검출 연구도 매우 빠르게 발전하는 분야이지만, 현재 가장 성공적인 모델로 평가받는 것은Faster R-CNN 모델이다. 이 모델의 특징은, 앞서 말한 물체가 있을만한 영역들을 열거하는 문제와 Feature Map을 학습하는 문제를 하나의 네트워크 구조로 동시에 풀어 정확도와 속도를 높였다는 점이다. 그 중, 얼굴을 검출해낸 것을 기반으로 얼굴 근육의 움직임을 분석해 감정까지 파악해낸다. 맨 눈으로는 알아보기 힘든 미세한 표정을 포착해 기쁨, 슬픔, 짜증 등 다양한 감정을 읽어 낸다.

* 1. **. 사진 설명 문장 생성 (Image/Video Captioning)**

딥러닝의 부상과 더불어 최근의 인공지능 연구의 특징 중 하나는, 서로 다른 인식 작업들을 하나의 네트워크 구조로 한꺼번에 풀기 위한 노력이 많이 진행되고 있다는 점이다. 특히 시각 인식과 자연어 처리 능력을 동시에 필요로 하는 문제들에 대하여 활발한 연구가 진행되고 있다. 그 중 대표적인 것은 사진 및 동영상 설명 문장 생성 (Image/Video captioning) 문제로, 사진 혹은 동영상 하나가 주어졌을 때, 이를 가장 잘 설명하는 하나의 자연어 문장을 만들어 내는 것이 목표이다. 이를 풀기 위한 딥러닝 모델은 주로 영상 정보를 표현하는 합성 곱 신경망 (Convolutional Neural Networks)과 연속적으로 단어를 생성하기 위한 순환 신경망 (Recurrent Neural Networks)이 결합된 형태로 구성된다. 이 분야에서 가장 많이 사용되는 모델은 주목 (Attention) 기반의 딥러닝 모델이다. 주목(Attention) 기법은, 인간이 사진 등을 보면서 시각 정보를 획득할 때 사진 상의 모든 세부 내용을 동일하게 보는 것이 아니라 시점을 바꿔가면서 사진 상의 다른 부분을 응시하여 이해한다는 점에 착안 한 방법이다. 즉, 사진 위의 서로 다른 부분을 집중하면서 단어 하나하나를 생성할 수 있도록 배우는 모델이다.

* 1. **. 자연어 처리 ( NLP )**

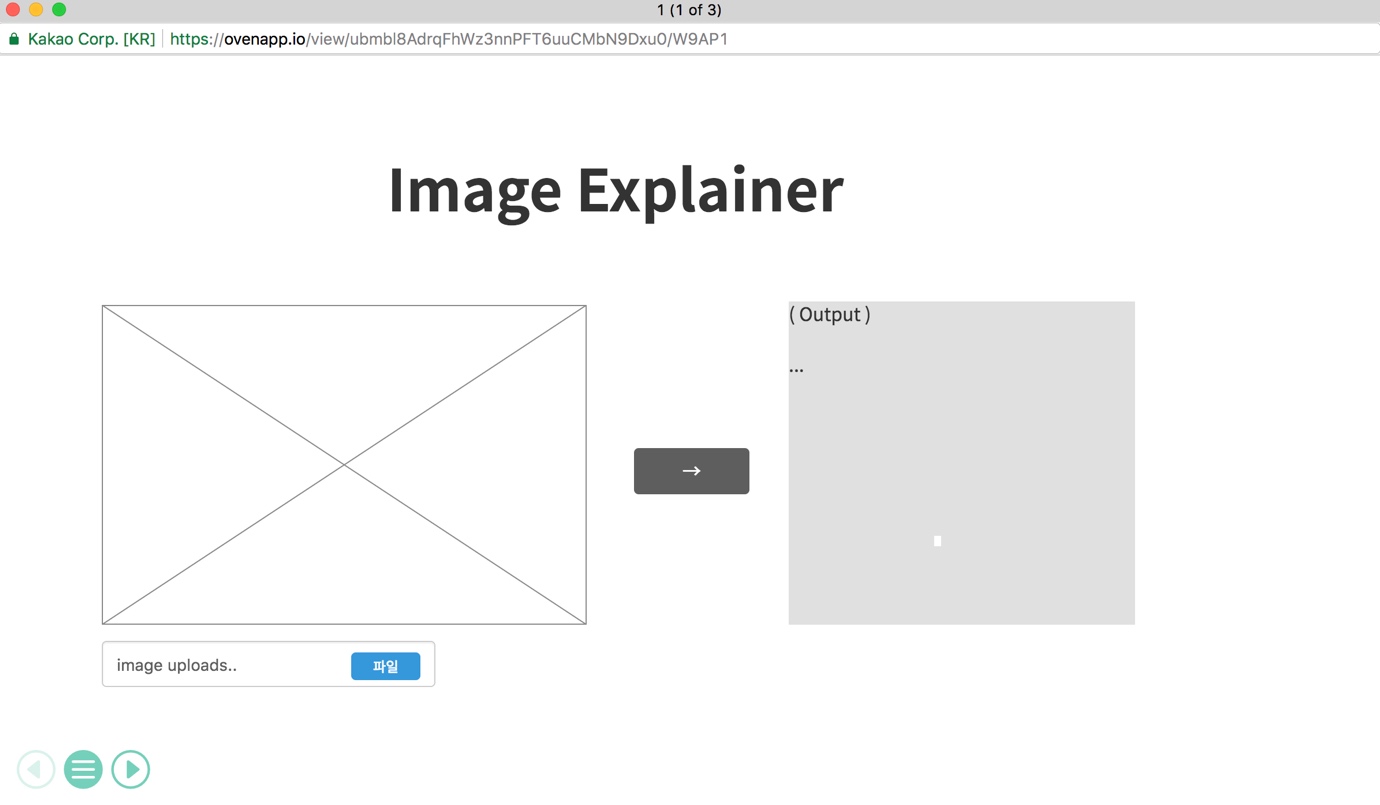
자연어란 인간이 일상적으로 사용하는 언어를 말한다. 반대되는 개념으로는 인공언어로서 컴퓨터가 처리하는 언어가 있다. 자연어 처리는 컴퓨터 과학, 인공지능, 언어학 이 3개의 분야가 합쳐진 분야이다. 자연어 처리를 하기 위해 컴퓨터가 자연어를 이해하는 과정은 다음과 같다.

1. 신호 처리( Signal Processing ) : a spoken bit of language 를 취해서 단어들의 순서로 변화시키는 작업으로 신호를 디지털화 하고 단어 조각 ( word segments ) 들을 구분한다. 취급되는 언어 요소는 phonemes ( 음소, 음성의 최소 단위 ) 이다.
2. 형태 분석( Morphological Analysis ) : 각 단어들이 그들의 구성요소로 분해되고, non-word tokens ( 구두점  같은 것 )은 단어로부터 분리된다. 예를 들면 "Bill's house" 에서 명사 Bill 과 접미사 's 로 구분된다. 취급되는 언어 요소는 morpheme ( 형태소, 뜻을 가지는 최소의 언어 단위 ) 이다.
3. 구문 분석( Syntactic Analysis ) : 단어들의 선형적 순서는 서로 간에 어떻게 관계되는지를 보여주는 구조로 변형된다. 이러한 파싱( Parsing ) 과정은 문장에서 단어들의 리스트를 "de-linearization" 시켜서 트리형식으로 변환한다. 문장의 syntax를 표현하기 위한 대표적인 문법에는 문맥 자유 문법 ( Context Free Grammar ) 와 확장천이문법 (Augmented Transition Network) 이 있다.
4. 의미 분석( Semantic Analysis ) : syntactic analyzer 에 의해 생성된 구조들에 의미 (meaning) 가 부여된다.  문장 "Colorless green ideas sleep furiously" (Noam Chomsky, 1957) 이 주어진다면 누구나 의미가 비정상 (semantically anomalous) 이라고 할 것이다. 이 단계에서는 각 단어가 지식베이스에서의 적절한 개체로 mapping 되어서, 각 단어의 의미가 서로 결합하는 방법으로 정확한 구조를 만들어야 한다. 의미구조를 표현하는 대표적인 문법에는 의미 망 (Semantic Network), 격 문법 (Case Grammar), 몬테규 문법 (Montague Grammar), 개념 의존 (Conceptual Dependency)이 있다.
5. 담화 통합( Discourse Integration ) : 각 문장의 의미는 앞선 문장에 의해 영향을 받고 다음 문장에 영향을 미친다. 문장에 포함된 존재들은 명확히 규정되어야 하고 이전의 존재와 관련되어 있어야 한다. 전체적인 discourse (이야기, 담화) 는 긴밀히 밀착 되어야 한다.
6. 화용 분석( Pragmatic Analysis ) : 말로써 표현하려던 것이 실제로 무엇을 의미하는 것이었는지를 결정하기위해 재해석되어야 한다.
   1. **. 기존 연구의 문제점 및 해결 방안**

현재 전문가들이 진행중인 연구는 일반적인 상황에서 진행을 하는 것이 대부분이다. 그런데 일반적인 상황, 현실 상황에 대한 이미지 데이터 셋은 거의 무한대로 존재하기 때문에 학습시간이 오래 걸릴 수 밖에 없다. 그리고 아직까지 정확도가 사용자가 익숙하게 사용하기엔 부족한 것이 사실이다. 네이버에서 최근 발표한 SmartSearch가 사용자에게 다가가려는 노력을 보이고 있고 광고도 많이 하긴 하지만 아직까지 사용자들이 사용하기엔 익숙하지 않은 면이 있다. 게다가 네이버의 SmartSearch는 단순한 물체검출의 기술 뿐이므로, 이미지를 단어가 아닌 문장으로 설명하는 소프트웨어가 필요하다.

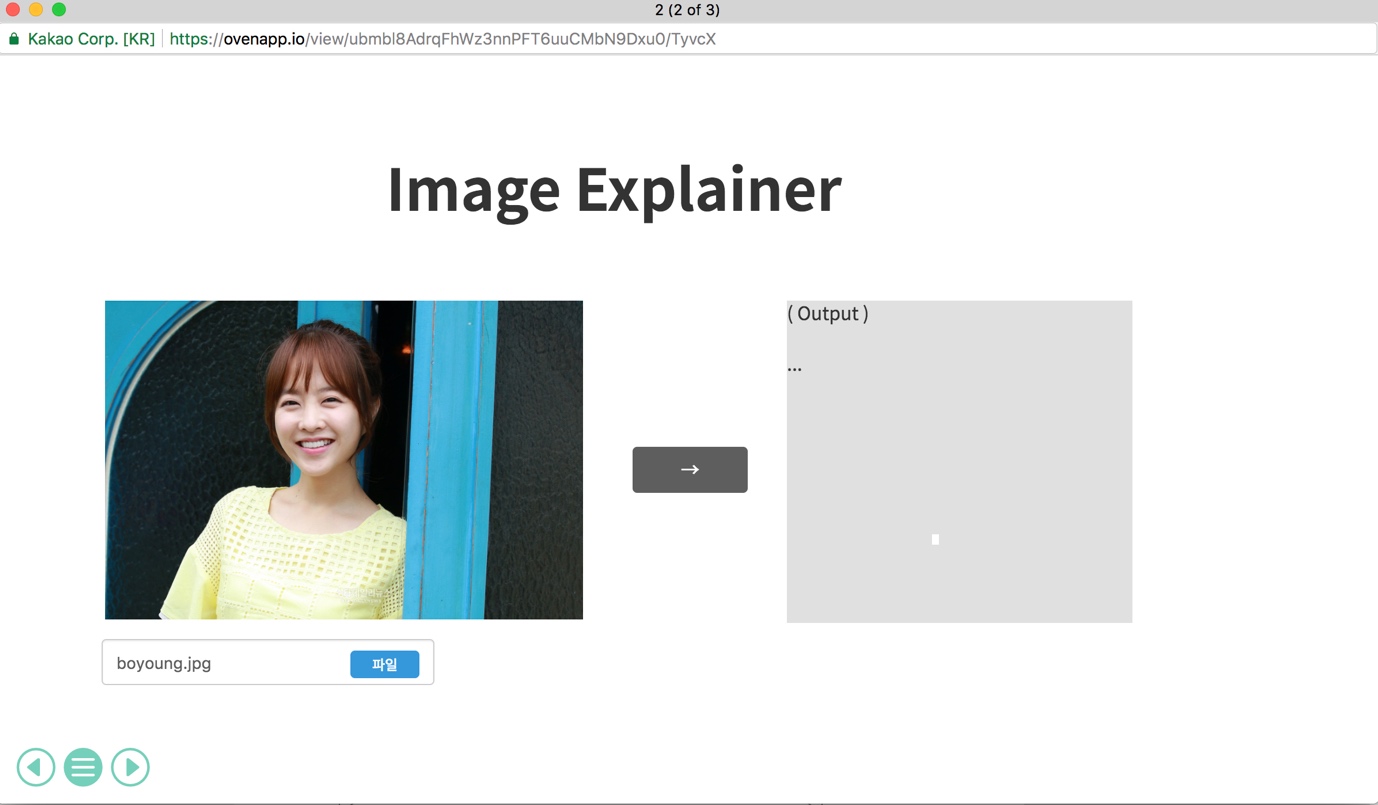
이를 해결하기 위해 캡스톤 디자인 1 프로젝트 정도의 규모에 적합한 상황을 찾아 특정한 상황에 최적화하여 정확도를 향상시키고 학습시간을 줄여보고자 한다.

1. **프로젝트 내용**
   1. **. 시나리오**
      1. **User Interfaces (UI)**



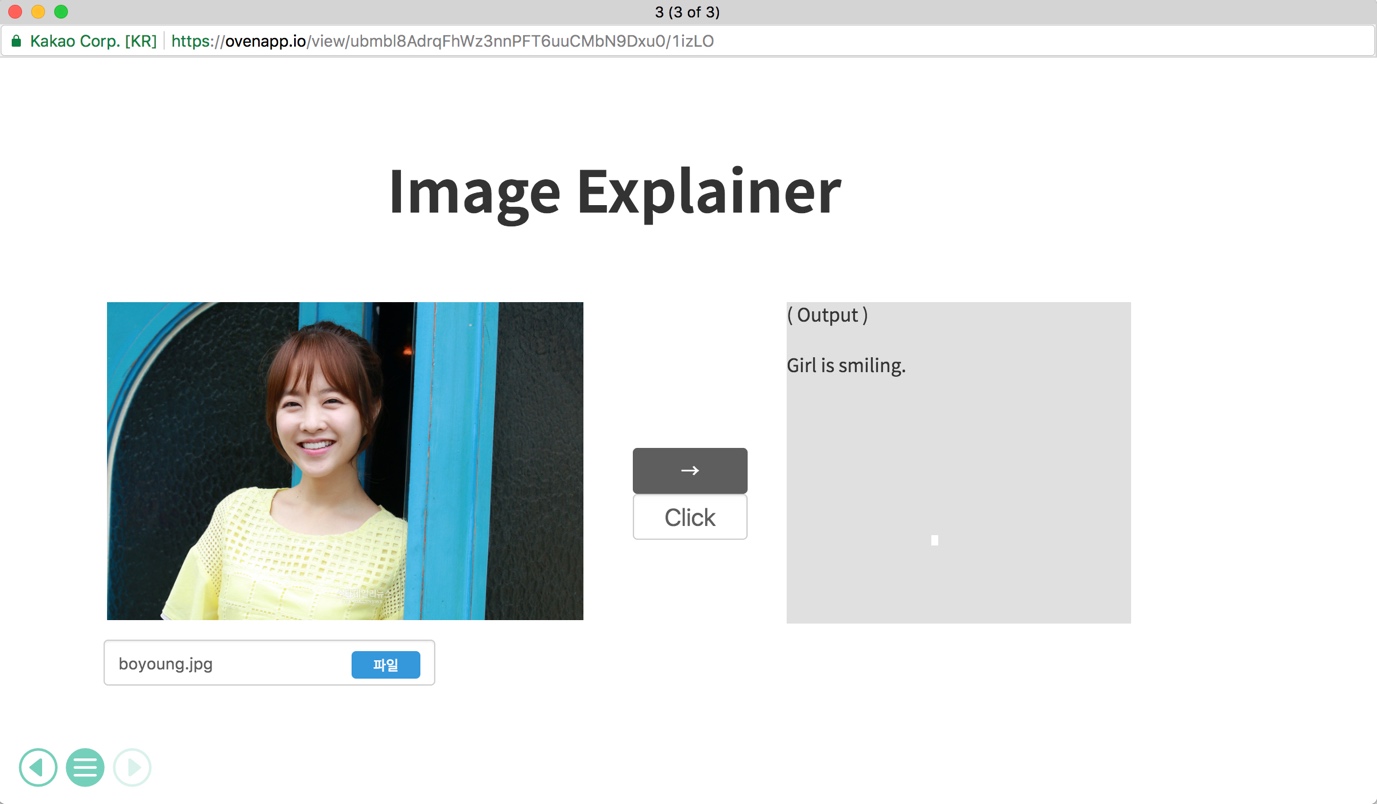
**[그림 1]**

**[그림 1]**은 ovenapp.io라는 UI Mock-up Tool을 사용하여 간단하게 디자인 해본 첫 페이지이다. 기본적으로 모든 사용자가 이용할 수 있으며 이미지를 업로드 하는 란과 이미지에 대한 설명이 Output으로 출력되는 란이 있다. 웹 페이지의 이름은 Image Explainer이다. 기본적으로 라우팅 없이 페이지 하나에서 이벤트 처리를 통해 동작하도록 할 생각이다.



**[그림 2]**

**[그림 2]**은 사용자가 boyoung.jpg 이미지 파일을 업로드 하고 클릭을 하기 전 대기 상태이다.



**[그림 3]**

**[그림 3]**은 사용자가 클릭을 통해 이미지에 대한 설명을 영문으로 얻는 것을 디자인 해보았다. 이미지에서 여자가 웃고 있고 그걸 텍스트 문장으로 표현한 output이 오른쪽에 출력된다.

* 1. **. 요구사항**
     1. **이미지 캡셔닝에 대한 요구사항**

원활한 이미지 처리를 위해 정해진 크기와 좋은 화질의 사진을 사용한다. 학습시킬 데이터셋은 Microsoft의 COCO 데이터 셋을 사용하고, 설정한 상황에 따라 필요한 데이터 셋을 추가적으로 만들어 사용한다. 이미지를 텍스트로 나타내기 위해서 사용할 수 있는 딥러닝 구조 중 RNN 구조와 LSTM 중 어떤 구조가 좋을지 연구해본다. 결과로는 사람이 이해할 수 있는 자연스러운 문장을 출력할 수 있도록 한다.

* + 1. **자연어 처리에 대한 요구사항**

데이터 셋과 성능 등을 고려해서 우리나라말인 한국어보다는 세계 공용어인 영어를 선택하였다. 이미지에서 뽑아져 나온 단어들을 분석하는 적절한 영어 형태소 분석기를 골라 전처리를 진행한다. 그리고 이미지에서 뽑아져 나온 단어들은 대부분이 물체나 사람, 감정 등 명사들일 것이다. 이 명사들을 어떻게 적절하게 동사를 섞어 가면서 적절한 문장을 만들 것인지가 관건이 될 것 같다. 지금까지 공부해 본 바로는 Word Embedding Vector로 단어들을 매핑하고 그걸 토대로 상황을 예측해서 문장을 만들어내는 것을 Tensor Flow를 통해 시도해보려고 한다. 그리고 자료가 많은 Python을 사용할 것이다.

1. **결론 : 기대효과 및 예상 일정**

기존에는 이미지를 텍스트로 설명하기 위해서는 사람의 힘이 필요했다. 하지만 딥러닝을 통해 사람의 힘을 거치지 않고 이미지를 설명할 수 있게 된다. 사진에 있는 물체를 검출하고, 사람을 검출하여 표정을 통해 감정을 인식한다. 그 후, 이미지에서 인식된 여러 물체들과 사람, 감정을 사람이 이해할 수 있는 자연스러운 문장으로 표현한다. 이러한 이미지를 텍스트로 설명해주는 기술을 시각장애인에게 가장 도움이 될 것으로 기대된다. 또한, 미술관의 사진이나 그림을 사람의 힘 없이 빠르게 설명할 수 있으며, 정지영상에서 동영상으로 발전시킨다면 TV 방송이나 영화를 설명해줄 수 있을 것으로 기대된다.

1. **참고문헌**

[1] <https://www.slideshare.net/ssuser06e0c5/i-64267027>

[2] <http://www.aistudy.co.kr/linguistics/natural/natural_language_processing.htm>

[3] <http://skccblog.tistory.com/3495>

[4] <http://solarisailab.com/archives/374>

[5] <http://mysnu.org/community/newtechnology.php?search_order=&search_part=&c_cate1=&mode=v&idx=10587&thisPageNum=>

[6] <https://pdollar.wordpress.com/2015/01/21/image-captioning/>