최종 발표

For

시각 장애인을 위한 상황 묘사 어플리케이션

Team 4

정준호(201113275) 전민규(201411802) 김도연(201614157)

Contents

프로젝트 소개

시각 장애인을 위한 상황 묘사 어플리케이션

주제 선정 목적

- 접근성이 좋은 스마트폰의 application을 사용하여 '시각장애인' 및 '저시력자' 등의 시각 보조
- 카메라로 찍히는 전방의 상황을 음성으로 알려줌

A man holding a child while standing at the fence of an elephant zoo enclosure.

The horse and puppy are separated by the mesh fence.

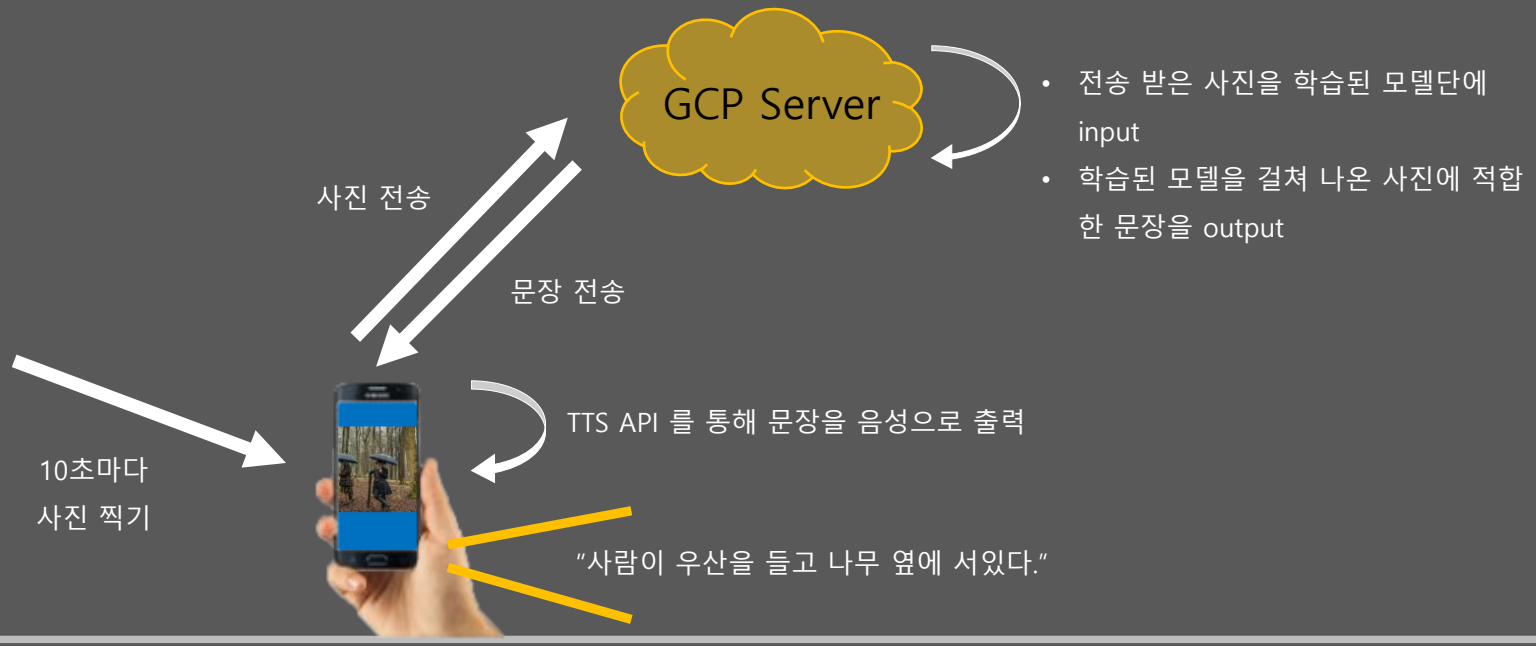
시각 장애인을 위한 상황 묘사 어플리케이션

' 시각 장애인을 위한 상황 묘사 어플리케이션 ' 이란 ?

시각 장애인분들을 위해 접근성이 쉬운 핸드폰 카메라로 찍히는 사진의 상황을 글 및 소리로 묘사해주는 어플리케이션입니다.

' 시각 장애인을 위한 상황 묘사 어플리케이션 ' 이 작동하는 방식은 ?

스마트폰을 들고 있으면 자동적으로 10초마다 한 장씩 사진을 찍어 그에 대한 설명을 음성으로 들려줍니다.



- 전송 받은 사진을 학습된 모델단에 input
- 학습된 모델을 걸쳐 나온 사진에 적합한 문장을 output

프로젝트 설계

요구사항 분석

기능 요구사항

- 전면부 상황 입력
  - 10초마다 주기적으로 사진 촬영
  - 촬영된 사진을 모델 단에 입력으로 넣음
- 딥러닝 모델 학습
  - MS COCO Captioning data 사용
  - 크게 모델은 Encoder , Decoder, Attention으로 구성
    - Image -> Encoder -> Decoder with Attention

요구사항 분석

기능 요구사항

- 학습된 모델을 통한 결과 도출
  - 학습된 모델을 사용하여 새로운 input image를 카메라를 통해 받았을 때 적절한 문장을 출력

- TTS
  - 학습된 모델로부터 나온 text 결과를 음성으로 변환하여 출력

요구사항 분석

비기능 요구사항

- 응답 시간 (성능)
  - 사진 한 장을 입력으로 넣었을 때 그 사진에 대한 적절한 설명이 텍스트 형태로 나오고, 음성 메시지로 들려주는 데 까지 10초

요구사항 분석

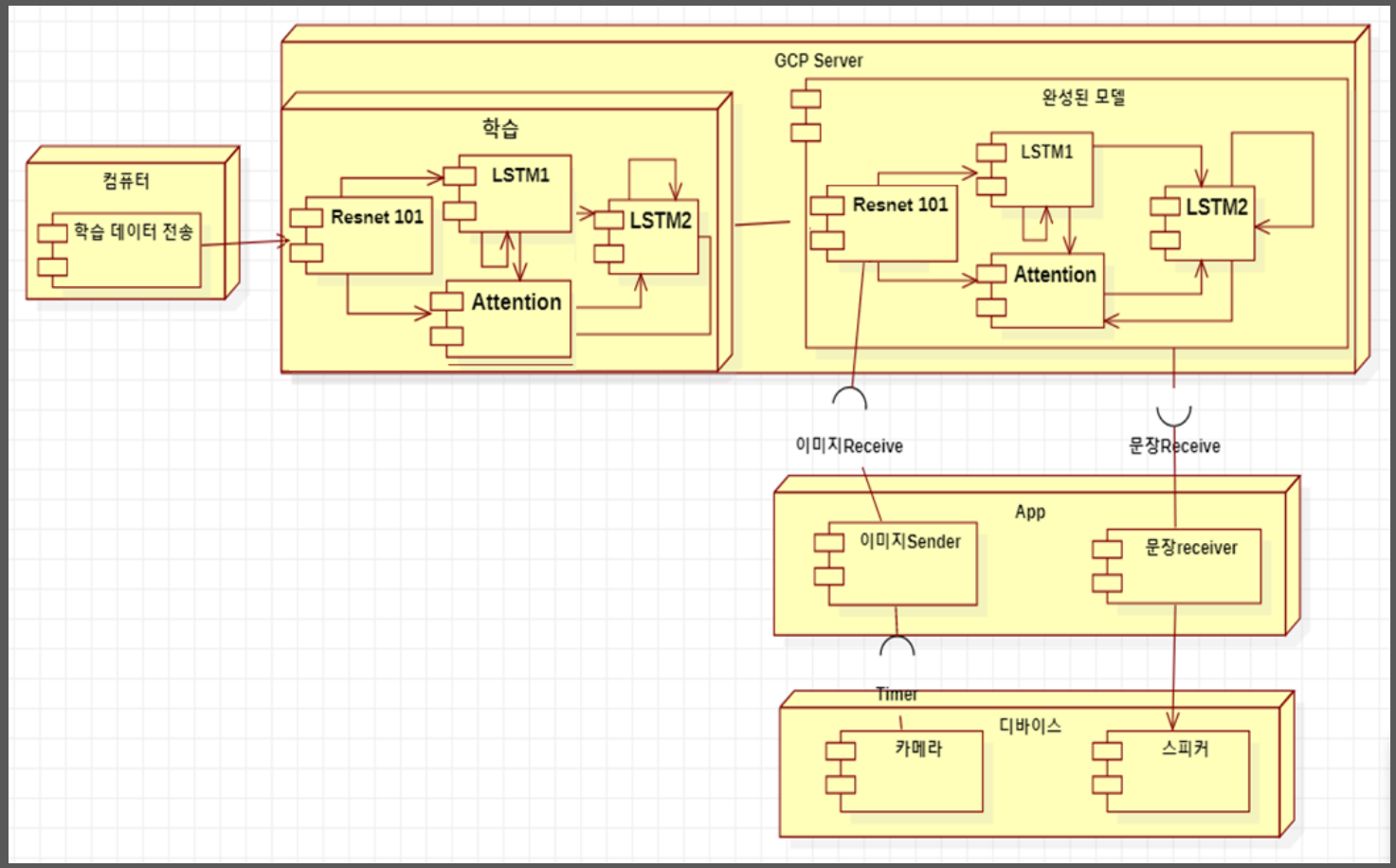
비기능 요구사항

- 상황 묘사 (신뢰성)
  - 다량의 train용 데이터로 학습된 모델을 사용하여 validation용 데이터에 대한 신뢰도 평가
    - BLEU-4 기준 0.251 이상인가 - 0.302
  - 다량의 train용 데이터로 학습된 모델을 사용하여 새로운 이미지에 대한 신뢰도 평가
    - 사람의 눈으로 봤을 경우 결과가 합리적인가

$$BLEU = min(1, \frac{output\ length(\text{예측 문장})}{reference\ length(\text{실제 문장})})(\prod_{i=1}^{4} precision_i)^{\frac{1}{4}}$$

Precision : n-gram을 통한 순서쌍들이 얼마나 겹치는가

Component Diagram



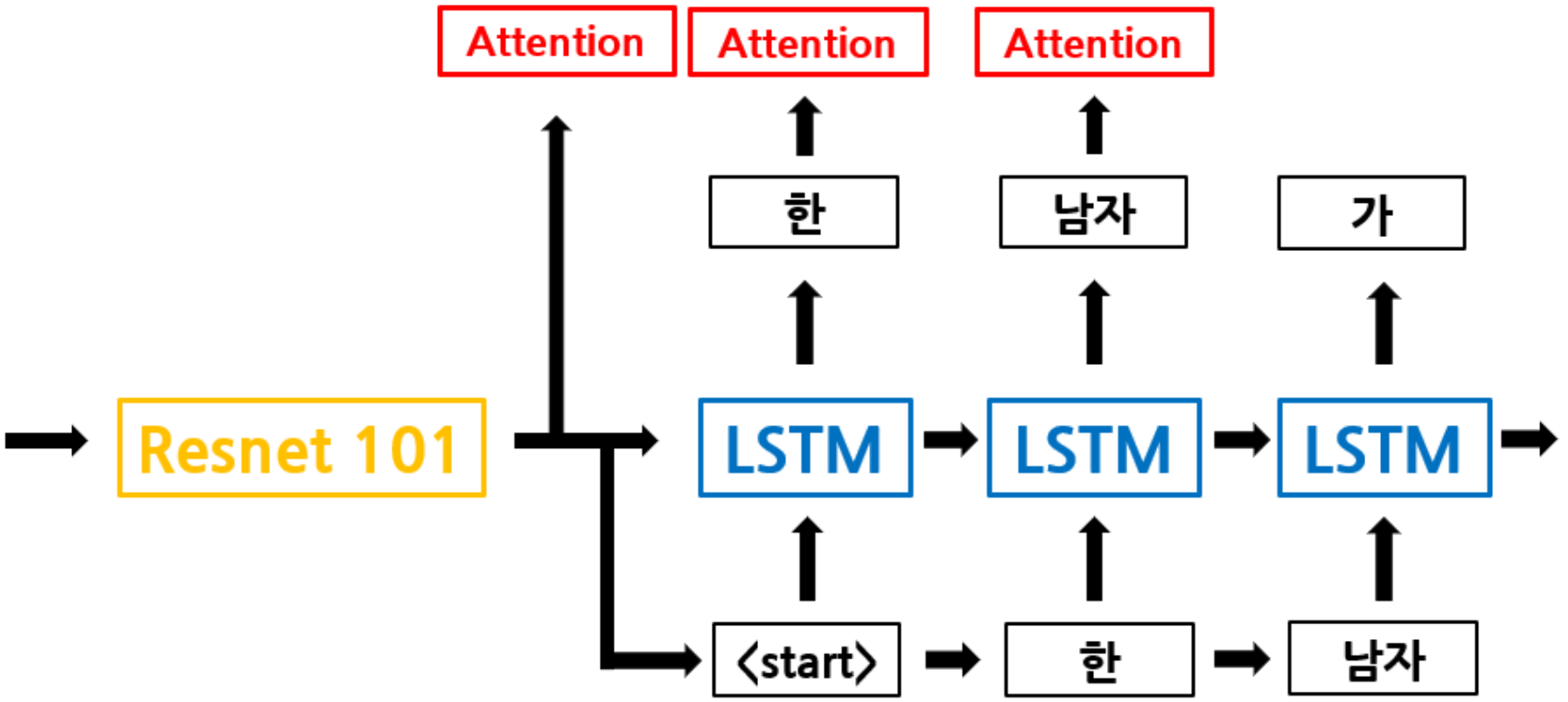
구현 내용

형태소 분석

<하늘을 나는 자동차>
Kkma
'하늘', '을', '날', '는', '자동차'
Komoran
'하늘', '을', '나', '는', '자동차'
Okt
'하늘', '을', '나', '는', '자동차'

<큰 건물 뒤에 서있는 흰색 자전거가 지나간다.>
Kkma
'크', 'ㄴ', '건물', '뒤', '에', '서', '어', '있', '는', '흰색', '자전거', '가', '지나가', 'ㄴ다'
Komoran
'크', 'ㄴ', '건물', '뒤', '에', '서', '어', '있', '는', '흰색', '자전거', '가', '지나가', 'ㄴ다'
Okt
'큰', '건물', '뒤', '에', '서있는', '흰색', '자전거', '가', '지나간다'

Model


Attention
Attention
Attention
한
남자
가
Resnet 101
LSTM
LSTM
LSTM
<start>
한
남자

Model

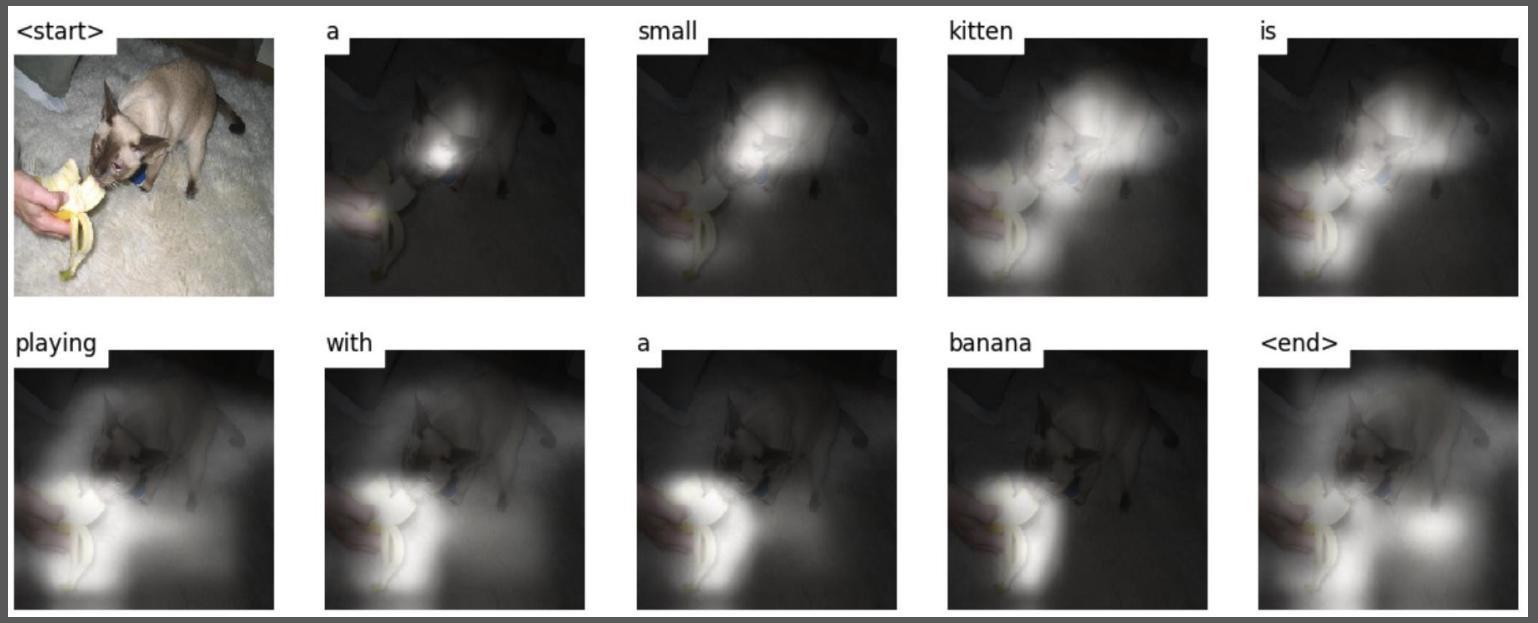
Image를 Input으로 받아 text를 output으로 출력하는 형태

- Image 인코딩 : Resnet101
  - 이미지 처리에 주로 쓰이는 Resnet을 사용
  - 이미지가 가진 공간정보를 유지하기 위해 Convolution feature map을 사용

- Image 디코딩 : LSTM1, 2
  - 문장이 Sequential한 형태로 구성되므로 LSTM을 사용
  - 인코딩 된 이미지를 보고 이미지 설명 문장을 생성

Model

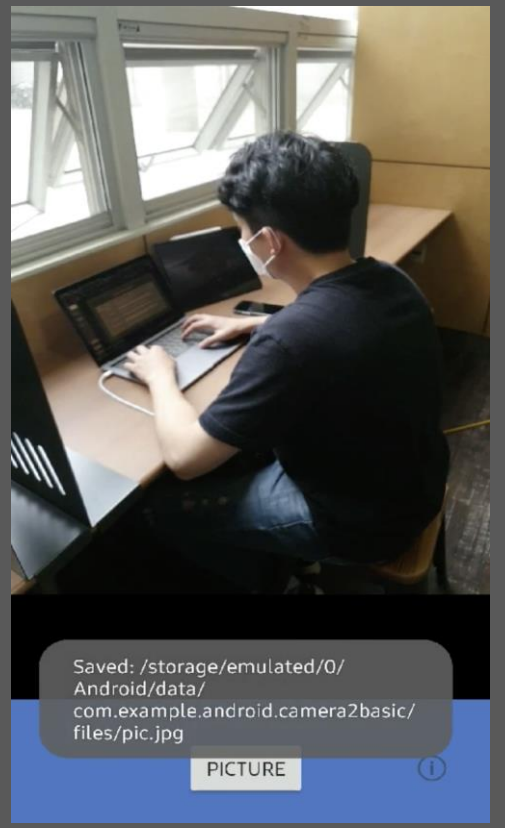
- Attention
  - 인코딩 된 이미지의 pixel값을 평균하여 바로 디코딩을 하는 것이 아니라, 어느 Pixel이 결과를 생성할 때 더 많은 정보를 주는 지를 학습을 통해 결정



Client (Android)

- 10초 마다 사진을 찍어 서버에 보내고 문장을 받는다.
- 0.8초 마다 새로운 문장이 왔는지 확인한다.
  - 새로운 문장이 기존 문장과 다를 때 tts를 호출한다.

```
2020-06-18 14:31:55.054 8695-8822/com.example.android.camera2basic I/System.out: 0.8초마다 음성 체크중
2020-06-18 14:31:55.279 8695-8695/com.example.android.camera2basic I/System.out: 10초마다 사진 찍기
2020-06-18 14:31:55.768 8695-8821/com.example.android.camera2basic I/System.out: 이미지 파일 크기 : 119858
2020-06-18 14:31:55.779 8695-8821/com.example.android.camera2basic I/System.out: 이미지 전송 완료
2020-06-18 14:31:55.779 8695-8821/com.example.android.camera2basic I/System.out: 결과 대기중
2020-06-18 14:31:55.855 8695-8822/com.example.android.camera2basic I/System.out: 0.8초마다 음성 체크중
2020-06-18 14:31:56.655 8695-8822/com.example.android.camera2basic I/System.out: 0.8초마다 음성 체크중
2020-06-18 14:31:57.456 8695-8822/com.example.android.camera2basic I/System.out: 0.8초마다 음성 체크중
2020-06-18 14:31:58.255 8695-8822/com.example.android.camera2basic I/System.out: 0.8초마다 음성 체크중
2020-06-18 14:31:59.056 8695-8822/com.example.android.camera2basic I/System.out: 0.8초마다 음성 체크중
2020-06-18 14:31:59.856 8695-8822/com.example.android.camera2basic I/System.out: 0.8초마다 음성 체크중
2020-06-18 14:32:00.656 8695-8822/com.example.android.camera2basic I/System.out: 0.8초마다 음성 체크중
2020-06-18 14:32:01.457 8695-8822/com.example.android.camera2basic I/System.out: 0.8초마다 음성 체크중
2020-06-18 14:32:02.257 8695-8822/com.example.android.camera2basic I/System.out: 0.8초마다 음성 체크중
2020-06-18 14:32:02.962 8695-8821/com.example.android.camera2basic I/System.out: 받은 문장 : 한 남자 가 노트북 을 사용 하면서 테이블 에 앉아 있다 .
2020-06-18 14:32:03.057 8695-8822/com.example.android.camera2basic I/System.out: 0.8초마다 음성 체크중
2020-06-18 14:32:03.859 8695-8822/com.example.android.camera2basic I/System.out: 0.8초마다 음성 체크중
2020-06-18 14:32:04.658 8695-8822/com.example.android.camera2basic I/System.out: 0.8초마다 음성 체크중
2020-06-18 14:32:05.284 8695-8695/com.example.android.camera2basic I/System.out: 10초마다 사진 찍기
2020-06-18 14:32:05.458 8695-8822/com.example.android.camera2basic I/System.out: 0.8초마다 음성 체크중
```



14:32:02.962 – 14:31:55.279 = 7.683

Testing

System Testing Results

TEST CASE	NAME	DESCRIPTION	Pass / Fail
1-1	전면부 상황 입력	• 스마트폰 앱에서 10초마다 사진이 찍히는지 테스트 한다.	Pass
1-2	전면부 상황 입력	• 새로 핸드폰 카메라로 '사람이 우산을 들고 나무 옆에 있는 사진'을 찍고 그 사진을 이미 GCP에서 학습된 weight를 갖고 있는 모델의 입력으로 넣고, 모델 입력 단에서 카메라로 찍은 사진 frame이 들어오는지 확인하는 코드를 추가한다. • Input image가 잘 들어오면 'OK', 아니면 'No Image'라는 메시지를 코드단에서 출력하도록 한다.	Pass
2	딥러닝 학습을 통한 모델	• Test data(train에서 사용하지 않은 MSCOCO 8000장)로 모델을 평가하였을 때, 평균적인 BLUE Score-4 기준 어절단위 0.111 의미형태소 단위 0.225 형태소 단위 0.251 이상이면 신뢰도를 갖고 있다고 판단한다.	Pass(0.302)
3	학습된 모델을 통한 결과 도출	• 위에서 정의한 '사람이 우산을 들고 나무 옆에 있는 사진' 과 그에 맞는 문장 '사람이 우산을 들고 나무 옆에 서있음' 을 준비해두고, 준비해둔 사진을 모델의 입력으로 넣었을 때, 미리 준비해둔 그 사진을 설명하는 문장 '사람이 우산을 들고 나무 옆에 서있음'과 유사한 문장이 10초 이내에 출력 되는지 확인한다.	Pass(7.683)
4	TTS (Text To Speech)	• 결과로 나온 문장을 TTS API를 통해 음성으로 바꿔서 출력하는지를 체크한다.	Pass

Pass / Fail Criteria

TEST CASE	NAME	Pass / Fail	Reason
1-1	전면부 상황 입력	Pass	➔ 10초마다 정상적으로 사진이 찍힌다.
1-2	전면부 상황 입력	Pass	➔ 앱과 서버를 연결시킴으로 앱이 전달한 사진이 완성된 모델단에 오류없이 input된다.
2	딥러닝 학습을 통한 모델	Pass	➔ 모델은 Encoder/Attention/Decoder 다 구현 Ex) Encoder : Resnet101 // Decoder : LSTM1 & LSTM2 구현 ➔ 일단 train 시켜본 결과, 정했던 BLUE Score-4 기준을 넘음
3	학습된 모델을 통한 결과 도출	Pass	➔ 앱과 서버를 연결시켜 앱이 디바이스의 카메라에게 사진을 요청한 순간부터 문장을 TTS API를 통해 음성으로 바꾼 뒤 디바이스의 스피커로 나오는 시간까지 10초안에 이루어지는 것을 확인할 수 있었다.
4	TTS (Text To Speech)	Pass	➔ 앱에서 그냥 임의의 문장을 출력하게 했을 때, 문장을 정확하게 읽었다.

Final Iteration 계획

TEST CASE	NAME	Pass / Fail	Plan
1-1	전면부 상황 입력	Pass	X
1-2	전면부 상황 입력	Pass	X
2	딥러닝 학습을 통한 모델	Pass	X
3	학습된 모델을 통한 결과 도출	Pass	X
4	TTS (Text To Speech)	Pass	X

졸업 작품은 끝이 났지만 전방의 상황을 좀 더 정확하게 출력할 수 있도록 계속해서 공부하겠습니다.

Q & A

감사합니다.