**산학 캡스톤 디자인1**

**Weekly Report #01 - 영상 음성 해설 웹서비스**

**팀명 : DD**

**컴퓨터공학과 이신우, 고훈현**

**1. 서론**

본 프로젝트는 시각장애인 또는 저 시력자를 위하여 영상 음성 해설 웹서비스 제공하여 실생활에 도움을 주는 것을 목표로 하고 있다.

**2. 기존연구**

**2.1 Image Captioning**

이미지 캡셔닝(Image Captioning)은 이미지를 설명하는 문장으로 변환시켜주는 기술로 인코더(encoder)를 이용하여 이미지의 특징을 추출하고, 추출된 특징들을 디코터(decoder)에 전달하여 워드 벡터 시퀀스를 출력으로 받는다. 출력된 시퀀스들은 단어사전과 매칭되며 최종적으로 사람이 알아볼 수 있는 문장으로 출력된다.

**2.2 Image Segmentation**

영상 분할(Image Segmentation)은 이미지의 각 픽셀이 어떤 클래스에 속하는지 분할하는 것이다. 영상 분할은 시멘틱 분할(Sementic Segmentation)과 인스턴스 분할(Instance Segmentation)으로 나눌 수 있는데, 전자는 픽셀 단위로 객체를 분류하는 것이고, 후자는 객체 탐지(object detection)를 추가하여 각 객체별로 분류한다.

**2.2.1 FCN**

FCN(Fully Convolutional Networks for Semantic Segmentation)은 시멘틱 분할을 위하여 네트워크의 모든 레이어를 컨볼루션 네트워크로 구성한 모델이다.

**2.2.2 SegNet**

SegNet은 자율주행에서 사용하는 영상 분할 모델이다. 모델은 인코더와 디코더가 결합된 형태이며, 인코더의 네트워크는 VGG16에서 fully connected layer를 제외한 컨볼루션 레이어를 그대로 사용한다. 디코더의 네트워크는 Upsampling과 컨볼루션 연산을 수행하며, 마지막 레이어에는 softmax 분류기가 배치되어있다.

**2.2.3 UNet**

UNet은 의료분야에서 영상 분할을 목적으로 설계된 것이다. 네트워크의 구성형태가 U자 형태를 띄고 있어서 UNet이라는 이름이 붙여졌다. 모델은 대칭구조로 구성되어 있으며, 수축 단계(Contracting path)와 팽창 단계(Expansive path)로 구성 되어있다. 전자는 입력의 맥락정보를 얻기 위하여 구성되었고, VGG를 기반으로 한 구조이다. 후자는 세밀한 Localization을 위하여 구성되었다.

**2.3 Visual Attention**

Visual Attention은 이미지의 중요한 부분에 집중하고 더 자세히 묘사하는 방법으로 Top-Down Approach와 Bottom-Up Approach의 장점을 합친 것이다.

**2.4 End-To-End Speech Recognition**

입력음성으로부터 하나의 통합 신경망을 거쳐 문자열 혹은 단어열을 인식하는 음성인식 방법을 뜻한다.

**2.4.1 Tacotron2**

Tacotron2는 구글에서 개발한 새로운 TTS모델로 텍스트를 입력받아 Spectogram을 생성하고 WaveNet Vocoder를 이용하여 음성합성을 제공하는 모델이다.

**2.4.2 Griffin Lim Vocoder**

Griffin Lim은 Spectogram을 음성으로 만들어주는 알고리즘이다.

**2.4.3 Wavenet**

Wavenet은 PixelCNN 구조를 기반으로 한 오디오 생성 모델이다.

**2.2 데이터셋 조사**

**2.2.1 MS COCO**

MS COCO 데이터셋은 Object Detection, Segmentation, Captioning에 대한 데이터를 약 330,000개를 보유하고 있는 데이터 셋이다.

**2.2.2 KSS**

KSS 데이터셋은 약 12시간 한국인 성우 음성파일과 대사로 이루어진 데이터셋이다.

**3. 제안 방법**

**3.1 문제 정의**

정지 영상에 대한 음성 해설을 만들고자 한다. 그리고 이것을 웹으로 서비스하는 것이 우리가 최종적으로 구현하고자하는 기능이다.

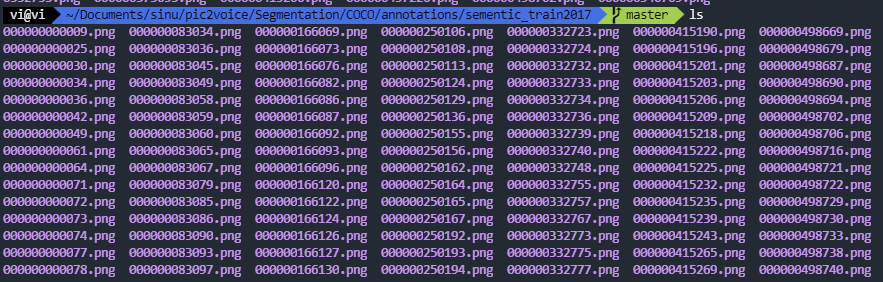
**3.2 자료 수집 및 선별**

**3.2.1 MS COCO 전처리**

****

**그림 1** 세그멘테이션 마스킹

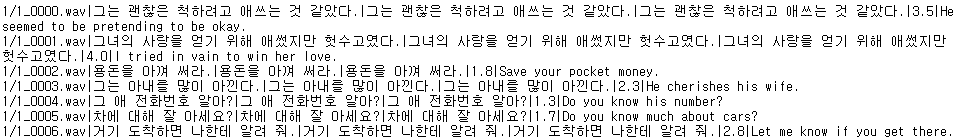
**(설명 추가)**

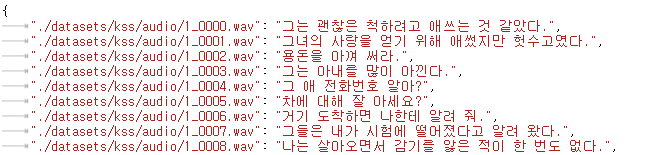
****

**그림 2** 전처리한 파일 목록

**(설명 추가)**

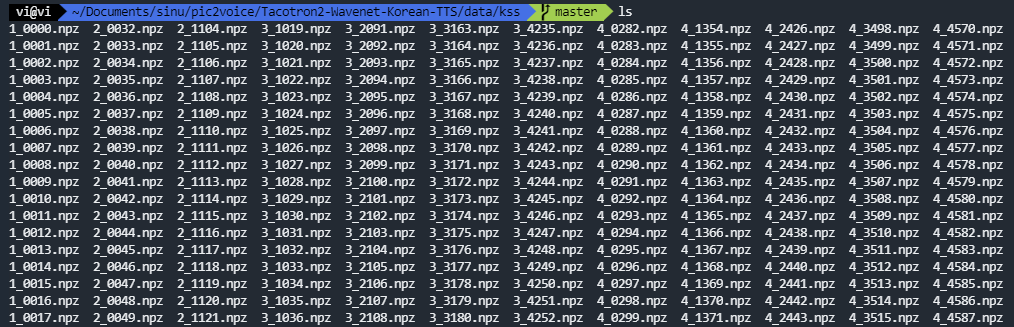
**3.2.2 KSS 데이터 전처리**

****

****

**그림 3** 스트립트 파일 json형식으로 변환

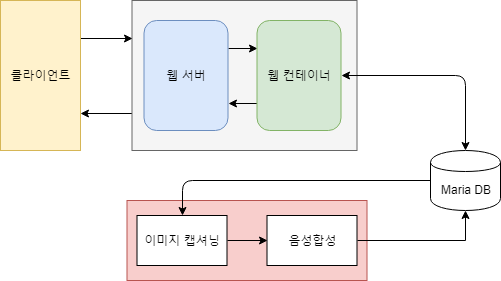
**(설명 추가)**

****

**그림 4** 전처리한 파일 목록

**(설명 추가)**

**3.3 전체 설계**

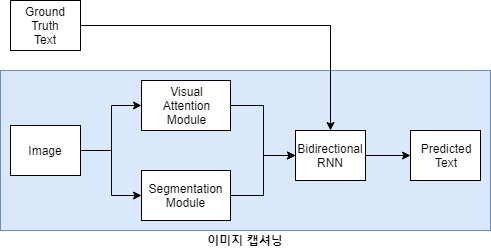
****

**그림 4** 전체 설계도

이미지를 음성으로 표현하는 프로그램을 만들기 위하여 이미지 캡셔닝 모듈과, 음성합성 모듈을 가지고 프로그램을 설계하였다. 그리고 웹서비스를 제공하기 위하여 웹서버, 웹컨테이너, DB가지고 설계하였다. 이미지 캡셔닝 모듈에서는 이미지를 입력받아 예측 문장으로 출력한다. 음성합성 모듈에서는 예측 문장을 받아 예측 음성으로 출력한다. 웹서비스는 클라이언트가 업로드 한 이미지를 DB에 업로드하고 이미지 캡셔닝 모듈과 음성합성 모듈을 거쳐 wav파일로 만든 뒤 DB에 저장한다. 이렇게 만든 wav파일은 웹 어플리케이션을 통하여 사용자에게 전달된다.

**3.4 세부 설계**

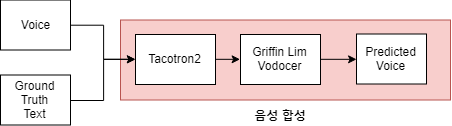
**3.4.1 이미지 캡션**

****

**그림 5** 이미지 캡셔닝 모듈

**(설명 추가)**

**3.4.2 음성 합성**

****

**그림 6** 음성 합성 모듈

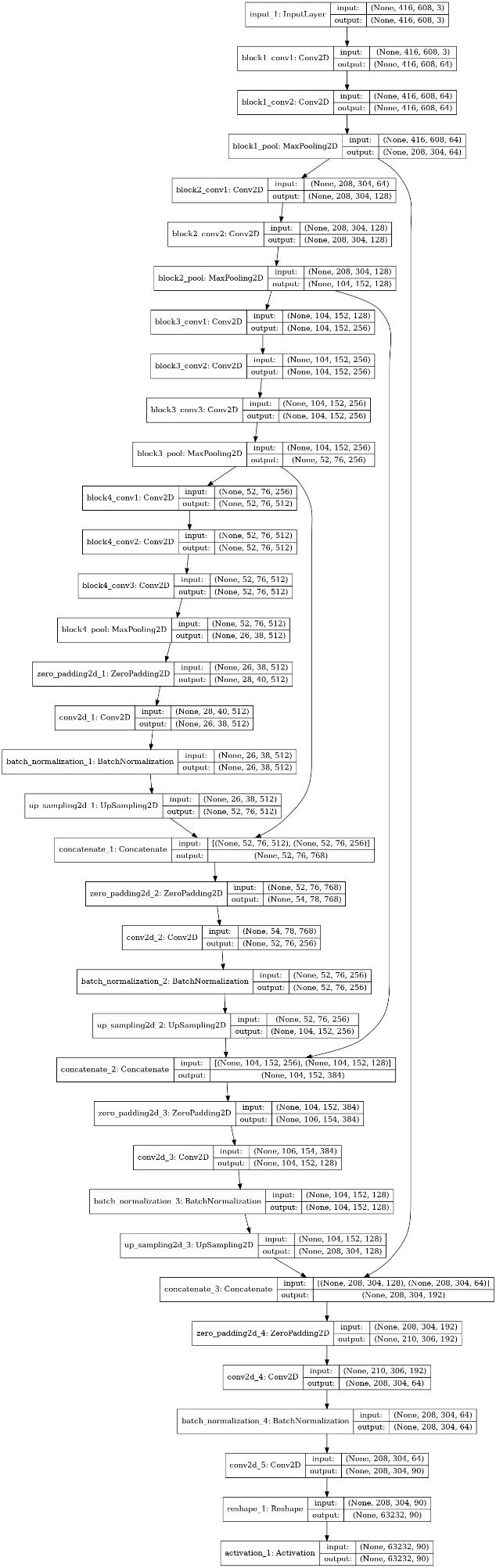
**(설명 추가)**

**3.4.3 웹 서비스 (추가 예정)**

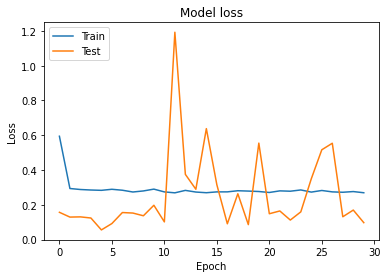
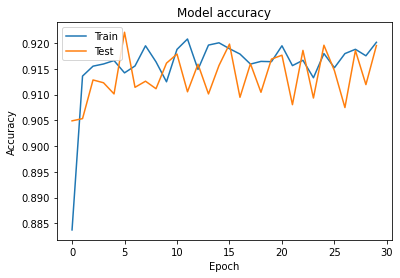
**4. 구현 및 실험**

**4.1 이미지 캡션**

**4.1.1 영상 분할**

****

**그림 7** 영상 분할 모델 구조

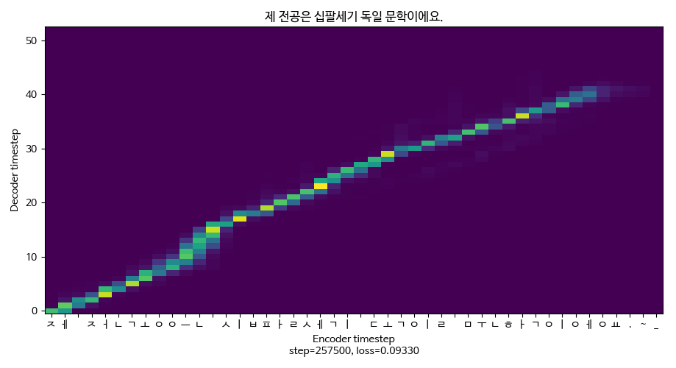
****

**그림 8** 영상 분할 모델의 정확도 및 손실 그래프

**(설명 추가)**

**4.1.2 이미지 캡션 (추가 예정)**

**4.2 음성 합성**

****

**그림 9** 타코트론2 음성 alignment 그래프

**(설명 추가)**

**4.3 웹 서비스 (추가 예정)**

**4.4 통합 테스트 (추가 예정)**

**5. 실험 및 결과**

**5.1 고찰 및 토론**

**6. 참고 문헌**

**(추가예정)**

**(설명 추가)**