**산학 캡스톤 디자인1**

**Mid Term Report**

**컴퓨터공학과 이신우, 고훈현**

**1. 관련연구**

**1.1 Panoptic Segmentation**

Panoptic Segmentation은 Semantic Segmentation(픽셀 별 레이블링)과 Instance Segmentation(물체를 인식하고 segment 하는 것)을 합친 개념이다. 이 개념에는 stuff와 thing을 구별하는 것도 포함하는 개념이다.

**1.2 Image Captioning**

Image Captioning은 이미지를 설명하는 문장을 만들어내는 것이다.

**1.3 Visual Attention**

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

**1.4 Tacotron2**

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

**1.5 LPCNet**

Mozilla에서 개발한 Vocoder로 DSP기술과 WaveRNN을 이용하여 신경망을 경량화하고 속도를 높인 모델이다.

**2. 프로그램 설계**

스크린샷이(가) 표시된 사진

자동 생성된 설명

**그림 1** 프로그램 설계도

이미지를 음성으로 표현하는 프로그램을 만들기 위하여 이미지 캡셔닝 모듈과, 음성합성 모듈을 가지고 프로그램을 설계하였다. 이미지 캡셔닝 모듈에서는 이미지를 입력받아 예측 문장으로 출력한다. 음성합성 모듈에서는 예측 문장을 받아 예측 음성으로 출력한다.

**3. 데이터셋 조사**

**3.1 MS COCO[[1]](#endnote-2)**

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

**3.2 KSS[[2]](#endnote-3)**

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

**4. 라이브러리 탐색**

**4.1 KoNLPy**

한국어 자연어 처리를 위한 파이썬 패키지이다.

**4.2 Mecab**

자연어 처리에서 사용되는 형태소 분석기이다.

**4.3 Google Cloud Speech-to-Text**

구글에서 제공하는 머신러닝 기반 음성 텍스트 변환 API이다.

**5. 진행상황**

음성합성 모듈의 Tacotron2 훈련이 진행중이고 그 중간결과는 **그림 2**와 **그림 3**에서 ground truth와 prediction의 mel spectrogram을 보여준다.

스크린샷, 컴퓨터이(가) 표시된 사진

자동 생성된 설명스크린샷, 시계, 컴퓨터이(가) 표시된 사진

자동 생성된 설명

**그림 2.** 좌: Step=300, 우: Step=25000

스크린샷, 시계이(가) 표시된 사진

자동 생성된 설명

**그림 3**. Step=52500

**6. 향후계획**

1. LPCNet Vocoder를 학습시킨다.
2. KSS를 이용하여 Tacotron2및 LPCNet을 학습시킨다.
3. 이미지 캡셔닝 모듈을 완성시킨다.
4. 실시간 이미지 음성 해설 웹서비스를 제공하기 위한 웹서버, 웹 어플리케이션, DB구축을 실시한다.

1. http://cocodataset.org/#download [↑](#endnote-ref-2)
2. https://www.kaggle.com/bryanpark/korean-single-speaker-speech-dataset [↑](#endnote-ref-3)