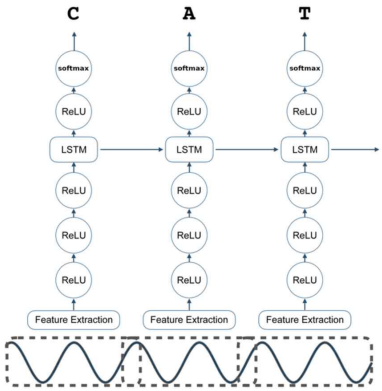
세종대학교

팀번호	23

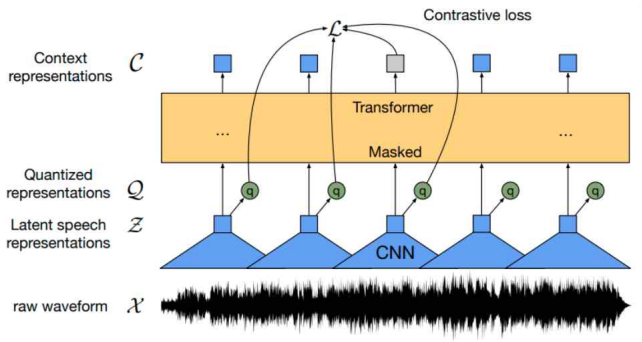
2024-1학기 창의학기제 주간학습보고서 (2주차)

창의과제	세종대학교 집현캠퍼스를 개선시킨 웹서비스 개발				
이름	박수진	학습기간	3/18~3/29		
학번	21011998	학습주차	2주차	학습시간	6
학과(전공)	인공지능학과	과목명	자기주도창의전공 I	수강학점	3
※ 수강학점에 따른 회차별 학습시간 및 10주차 이상 학습 준수					
금주 학습목표	음성 데이터를 텍스트로 바꾸는 방법인 STT 모델에 대해 익히고 직접 테스트하여 음성 데이터를 텍스트로 변환한다.				
학습내용	1주차에 저장한 오디오 파일을 텍스트로 변환하여 저장하는 과정을 진행하였다. 소리 데이터를 텍스트로 변환하기 위해서는 음성을 인식하고 이를 텍스트로 변환하는 STT(Speech-to-Text) 모델을 사용할 수 있다. 그래서 우선 이러한 STT 파이썬 모델과 라이브러리 등을 조사하였다. Clova Speech Recognition, Google Cloud Speech-to-Text, Whisper, Speech Recognition, Mozilla의 DeepSpeech, wav2vec 등이 이에 해당하는 것들이다. - Clova Speech Recognition : 네이버 클라우드 플랫폼의 서비스이다. 비로그인 오픈 API를 제공한다. 다양한 기능들을 제공하고 사용방법 또한 자세하게 나와있다. - Google Cloud Speech-to-Text : Google Cloud Platform에서 API를 제공한다. 기계학습과 음성처리 기술을 기반으로 한다. 그리고 다양한 언어를 지원하기 때문에 다양하게 활용할 수 있다. - Whisper : OpenAI에서 개발한 인공지능 모델이다. 자동 음성 인식 및 음성 번역을 위해 사전 훈련이 되었다. 소리를 텍스트로 변환하는 오픈 소스 STT(Speech-to-Text) 모델이다. github에 코드가 공개되어있어 쉽게 사용이 가능하다. - Speech Recognition : Python에서 라이브러리를 설치하고 import 해서 사용할 수 있다. - Mozilla의 DeepSpeech : Mozila가 개발한 오픈소스 음성인식 엔진이다. 기계학습 엔진으로 되어있으며 딥러닝 알고리즘을 사용한다. TensorFlow를 기반으로 만들어졌고 파이썬의 경우 패키지를 통해 이용 가능하다. 또한 DeepSpeech는 장단기 메모리(Long Short-Term Memory, LSTM) 네트워크를 기반으로 한다. LSTM은 시간적인 패턴을 파악하고 기억하는 능력이 뛰어나서 시간적 순서가 중요한 음성데이터를 처리할 때 적합하다. - Wav2Vec : 자동 음성 인식(ASR)을 위해 사전 훈련된 모델이다. self-supervised learning(자기주도학습)을 기반으로 한다. self-supervised learning은 라벨이 없는 데이터를 이용하여 자기 자신의 특성(representation)을 학습하는 방법이다. Wav2Vec 라벨이 없는 50,000시간 이상의 음성 데이터를 학습하였다. 이렇게 사전에 대량의 데이터로 학습을 한 후 적은 양의 데이터만을 이용해 미세조정(Fine-tuning)을 거쳐 사용할 수 있다.				

<DeepSpeech 동작 원리>



<Wav2Vec framework>



그 후 이 중 몇가지를 사용하여 소리를 텍스트로 변화하는 것을 시도해보았다.

가장 먼저 DeepSpeach 모델을 시도해 보고자 하였다. 그런데 DeepSpeach는 사전 학습 모델이 아니기 때문에 이 모델을 학습시키기 위해서는 대량의 레이블 된 음성 데이터가 필요했다. 또한 대량의 데이터를 학습시키기 위해서는 시간과 컴퓨터 자원이 많이 필요한 문제가 있어 현재 상황에서는 이 모델을 사용하기에 어려움이 있다.

다음으로 시도한 모델은 Whisper이다. github에 공개된 코드를 찾아서 직접 실행했다. 오디오 데이터를 입력해 실행해 본 결과, 입력 데이터의 길이 제한이 발생했다. 1시간 분량의 데이터를 입력해도 30초 분량까지만 텍스트로 변환된 결과가 나왔다. 길이가 30초 정도인 짧은 데이터만 가능한 것이다. 또한 이 모델은 영어로 된 데이터에 특화되어 있어 한국어 데이터에 대해서 성능은 별로 좋지 않았다.

다음으로는 Google Cloud Speech-to-Text를 시도해 보았다. Google Cloud API를 발급받아서 사용하면 되기 때문에 사용하는 것은 어렵지 않았다. 그리고 데이터를 넣어서 테스트를 했는데 결과도 잘 나온 것을 확인했다. 그러나 데이터의 길이가 한 달에 60분을 넘어갈 경우 비용이 발생하는 문제가 있어 다른 방법을 사용하기로 했다.

다음은 파이썬 라이브러리인 speech recognition를 사용했다. 그러나 이 방법도 문제가 있었다. 짧은 데이터에서는 잘 작동을 하나 데이터의 길이가 길어질 경우 오류가 발생했다. 그

<table>
<tr><td></td><td>래서 10분, 5분, 4분, 3분 분량의 데이터로 각각 테스트해 본 결과 최대 4분의 데이터까지 가능한 것을 알 수 있었다. speech recognition도 whisper와 마찬가지로 입력 데이터 길이의 제약이 있지만 whisper를 사용했을 때보다 더 정확하게 텍스트로 변환되었다. 그래서 입력할 오디오 데이터를 4분 단위로 자른 후 speech recognition을 통해 텍스트로 변환하고 그 결과들을 다시 합치는 방식을 선택하였다.

그런데 라이브러리를 가져와서 사용하는 것도 좋지만 직접 모델을 훈련시켜 사용하는 것도 의미 있을 것으로 생각한다. Wav2Vec은 사전 학습 모델이기 때문에 적은 양의 레이블이 지정된 음성 데이터만 있어도 미세 조정(fine tuning)을 통해 사용이 가능하다는 장점이 있다. 그래서 Wav2Vec 모델도 사용해 보고자 한다.</td></tr>
<tr><td>학습방법</td><td>구글 검색을 통해 여러 가지 STT모델 및 라이브러리를 조사하였다. 잘 모르는 개념이 나왔을 경우 추가 검색을 통해 해결하거나 책을 활용하여 해당 개념에 대해 학습하였다. github에 공개된 코드를 참고하여 코랩에서 직접 코드를 실행하며 코드에 대한 이해도를 높이고자 하였다.</td></tr>
<tr><td>학습성과
및
목표달성도</td><td>• 이전 주차에 추출한 오디오 데이터를 텍스트로 변환함
<변환된 텍스트>
소리지 우리 교재는 그림이 없지만 선생님 이렇게 설명
라는 허상열 또 얘기하는 거야 일본 같은 거 알겠지 안
유동성이 있는지 유동성이 있기 때문에 움직일 수가 있
책도 이런 철주 분석 같은 성분으로 되어 있겠구나라고
했다 팽창을 해지면서 그 뒤에 있는 공기를 미는 겁니다
게 중요하지 않아요 거의 없어요 제일 나중에 편협하다
하는 방법과 지난까지의 거리를 구하는 방법이 달라요
워 도착할 때까지 요거 가지고 알아내는 거야 요거 재다</td></tr>
<tr><td>참고자료
및 문헌</td><td>Whisper 참고 자료
https://learn.microsoft.com/ko-kr/azure/ai-services/speech-service/whisper-overview
https://github.com/openai/whisper
SpeechRecognition 참고자료
https://pypi.org/project/SpeechRecognition/</td></tr>
<tr><td>내주 계획</td><td>아직 실행해 보지 못한 wav2vec 모델을 fine tuning해서 사용해 본다. 그리고 speech recognition을 사용해서 나온 결과와 wav2vec을 강의데이터에 맞게 fine tuning한 모델을</td></tr>
</table>

세종대학교

	통해 나온 결과를 비교한다.

년 월 일

지도교수 (인)