

Lip2Text

데이터청년캠퍼스 5조

송문영 안혜준 이우준 이태희 임준혁

01

주제 선정

주제 선정 배경
사용할 데이터

02

프로젝트 진행과정

단어 단위 기반 학습
문장 단위 기반 학습

03

실제 적용 및 가능성

커스텀 데이터셋
한계점 및 활용방안

01

주제선정

주제 선정 배경
사용할 데이터

02

프로젝트 진행과정

단어 단위 기반 학습
문장 단위 기반 학습

03

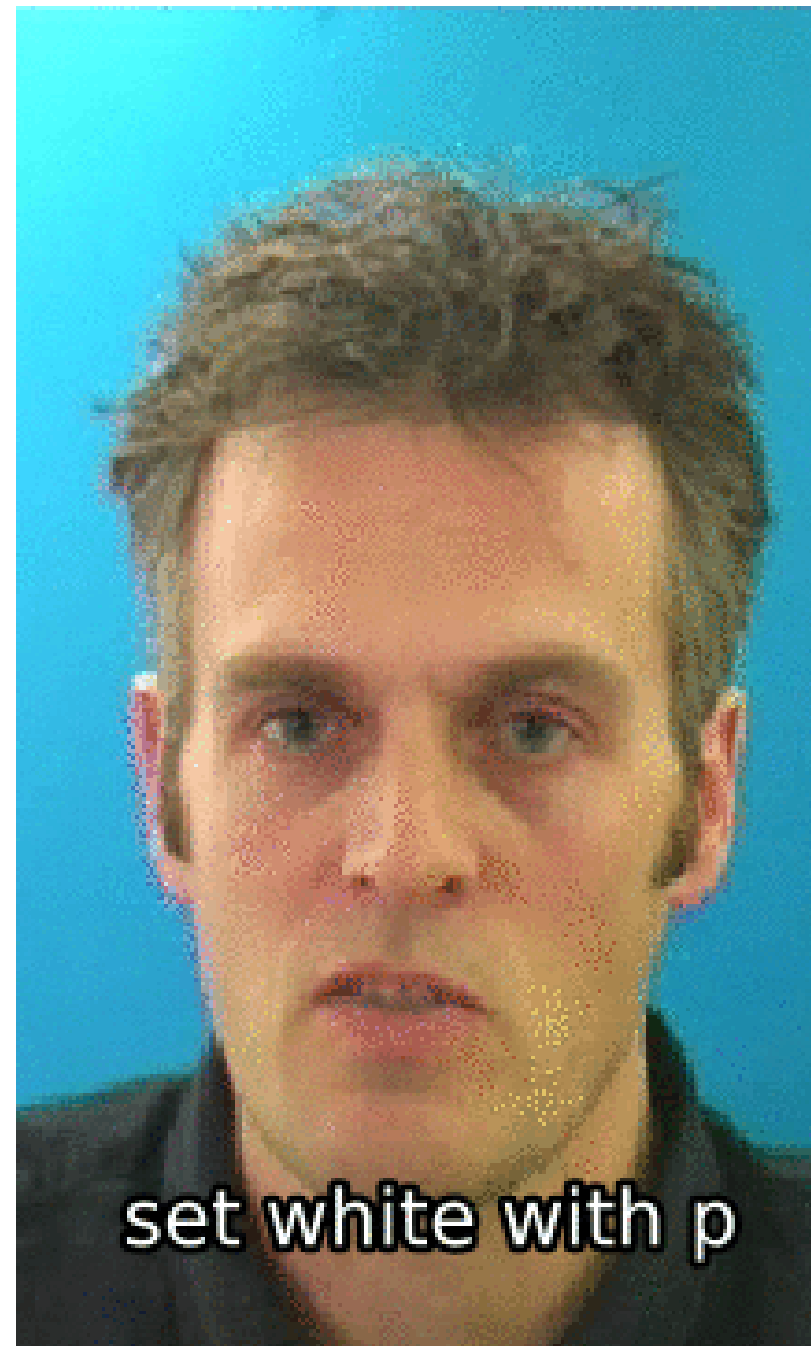
실제 적용 및 가능성

커스텀 데이터셋
한계점 및 활용방안

LipReading

한글 독순술 모델

딥러닝과 자연어처리를 이용한
한국어 독순술 모델 제작



- 01 청각장애인들의 일상생활에서 불편함
오랜 시간에 걸쳐야 독순술이 습득이 가능
- 02 잡음이 많이 생기는 화상회의
음성이 제한되는 상황에서 적용 가능성

모델 학습 데이터

단일 화자가 단어 및 문장에 대해
발음하는 영상 및 자막



LRS2(Lip Reading Sentences 2) Dataset

'부족한 데이터셋'

- + AIHub에서 제공하는 데이터 셋
자주 사용하는 한글 단어 1000개에 대해 발음하는 영상 제공
- 영어와 중국어의 경우는 LRW, LRW1000 등 다양한 데이터 셋 존재
BBC의 방송에서 발음하는 단어 단위로 잘라내 처리되었음

한국어에 대한 데이터셋은 AIHub가 유일



한국어 데이터셋

AIHub 및
TV 뉴스 동영상

MBC, KBS, SBS 등 다양한 방송사의
8년간의 뉴스 영상을 수집함 (대략 30만개)

OpenCV와 ffmpeg 등 이미지 처리 모듈들을
활용해 발음하는 사람의 얼굴을 동영상으로 처리함.

01

주제선정

주제 선정 배경
사용할 데이터

02

프로젝트 진행과정

단어 단위 기반 학습
문장 단위 기반 학습

03

실제 적용 및 가능성

커스텀 데이터셋
한계점 및 활용방안

LipReading

프론트엔드와 백엔드 모듈에 각각 다른 모델을 이용
모듈 각각은 local motion(단어, 프레임) / sequence 레벨의 패턴에 집중한다

01

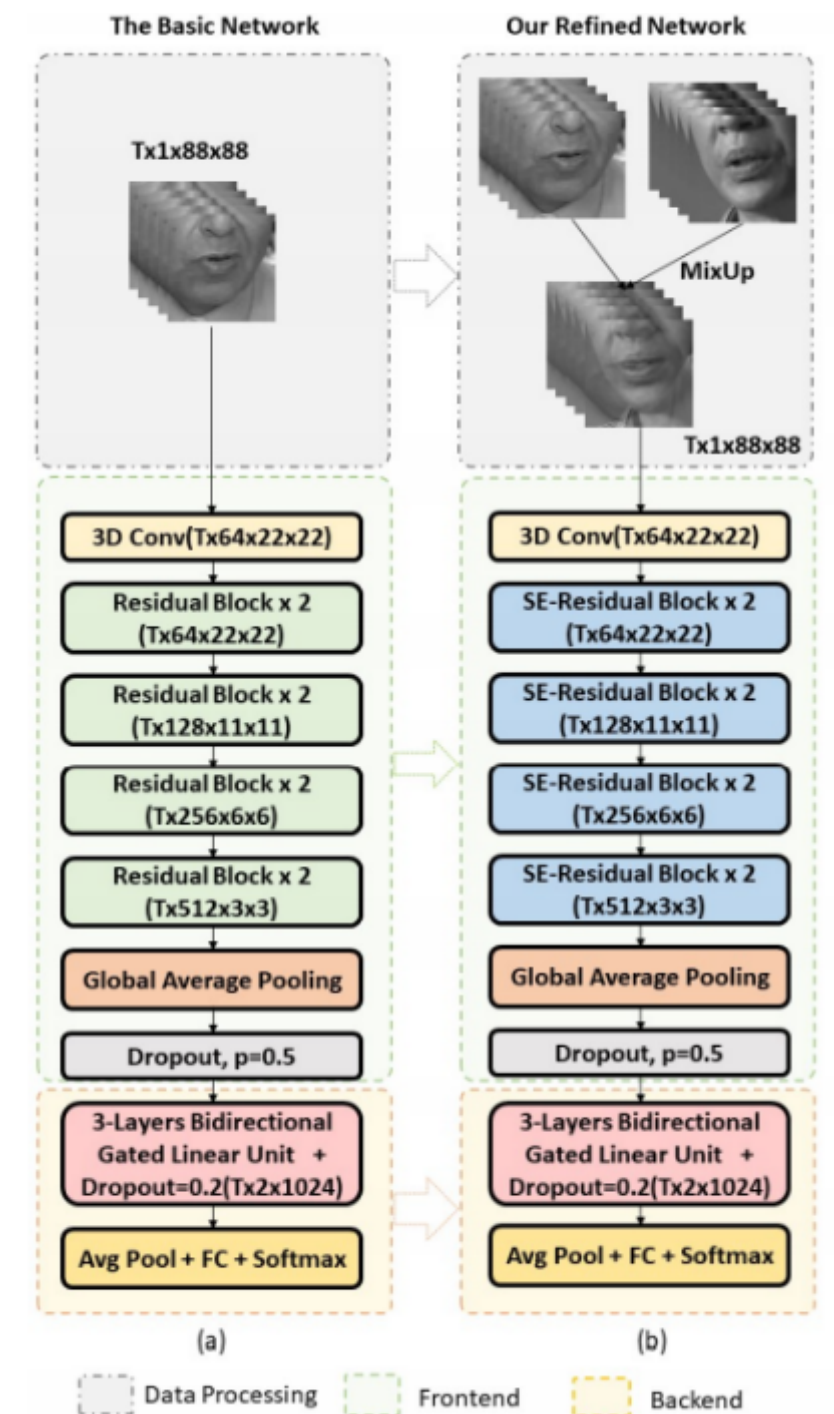
프론트엔드 모듈에는 ResNet-18,
백엔드 모듈에는 GRU를 이용

03

PyTurboJPEG, OpenCV 등을 이용한
전처리 후 학습 진행

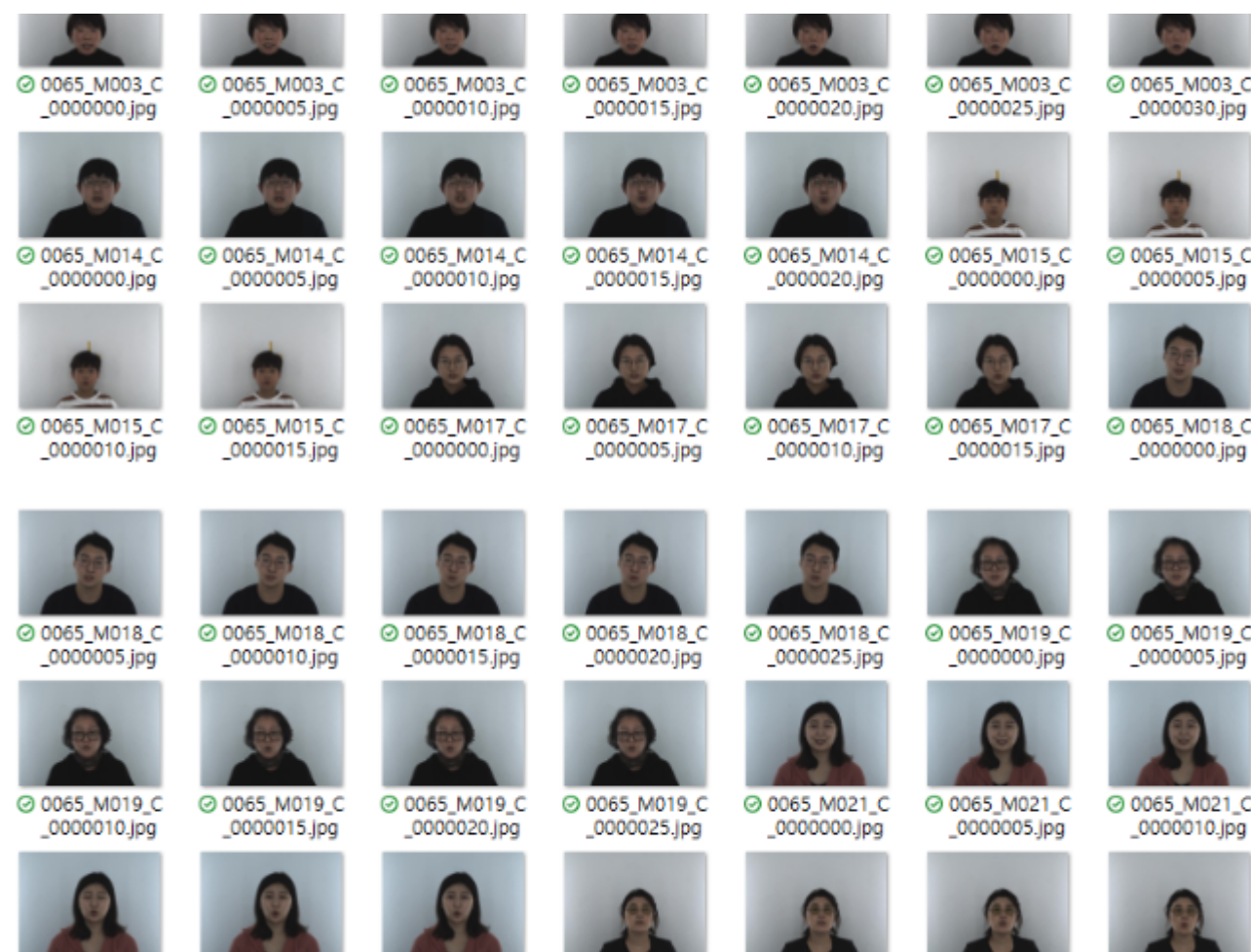
02

단어별로 분류되어 있는 한국어 독순술
데이터셋 '신체 말단 움직임 영상' 이용



공개된 데이터의 한계

설명서에는 1000개의 단어와 50만개의 영상이 제공된다 했으나,
실제로는 베타 버전으로 감탄사/대명사 등 극히 일부의 프레임 사진만을 제공



한계점

AIHub의 베타버전

10%

1000개 어휘 중 양이 적은 감탄사, 관형사 등만 공개됨

미정

데이터가 아직 처리중이고, 공개 일정이 미정

네이버 뉴스 데이터

크롤링한 30만개의 영상

Alignment

영상과 대본의 싱크를 맞추는 문제

단어 단위

실제 alignment를 단어 단위로 처리할 시 너무 많은 시간이 걸릴 것

문장 단위 학습

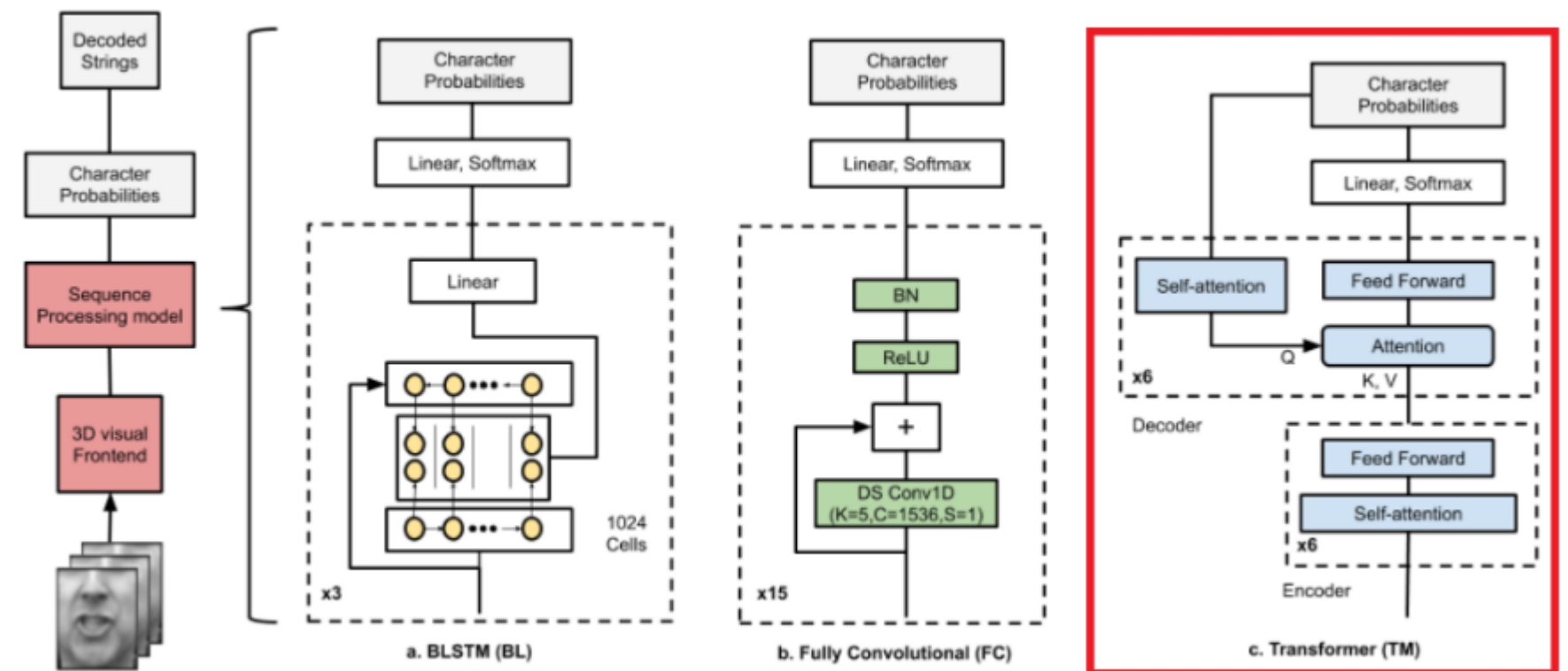
일부의 논문에서 메인 모델의 학습 및 예측을 단어가 아닌 문장 단위로 진행
이를 이용해 모델을 학습하고자 시도함

문장 단위 Alignment

유튜브에서 생성되는 자동자
막을 이용해 싱크를 맞춤

Deep Lip Reading

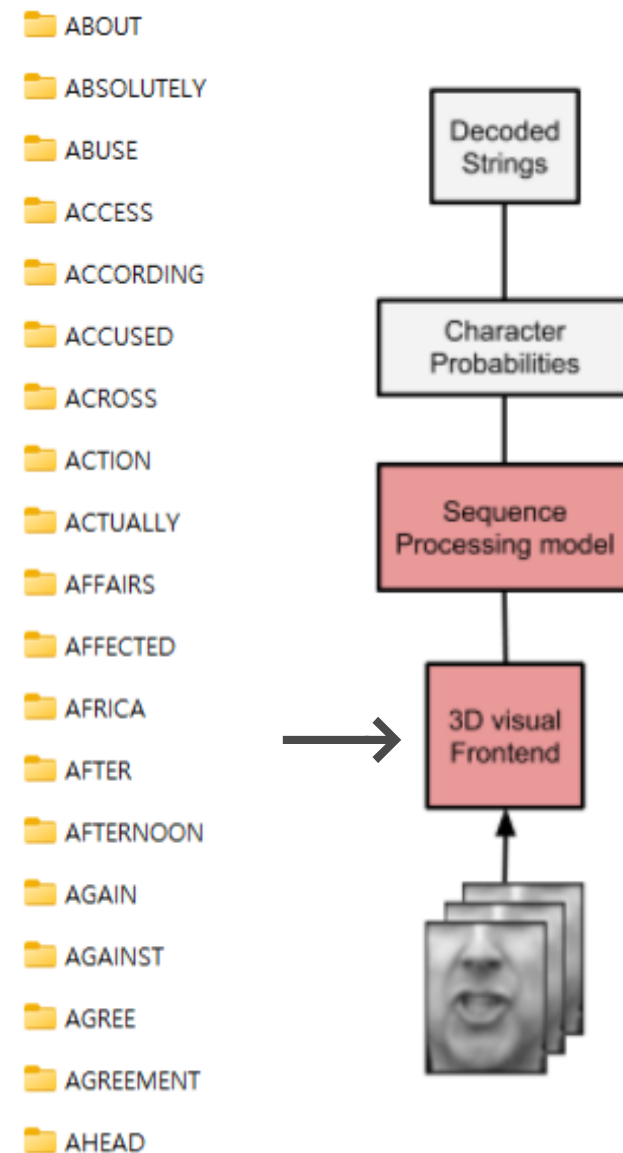
문장 단위



STCNN + Transformer (+ Language Model)

단어 단위 분류의 필요

맨 처음 Visual Module에서 단어 별 분류된 데이터셋으로 Pretrain이 필요함을 너무 늦게 깨달음



LRW : 500 words

단어 단위 전처리

전처리 시간을 줄이기 위해 문장 단위로 전처리를 했으나, 결국 단어로 Label된 데이터 세트의 필요성

영어 모델 이용

한국어로 된 단어 데이터 세트 존재하지 않으므로, LRW로 pretrain 된 가중치와 모델을 이용해 논문과 다른 데이터셋에 적용을 시도

01

주제선정

주제 선정 배경
사용할 데이터

02

프로젝트 진행과정

단어 단위 기반 학습
문장 단위 기반 학습

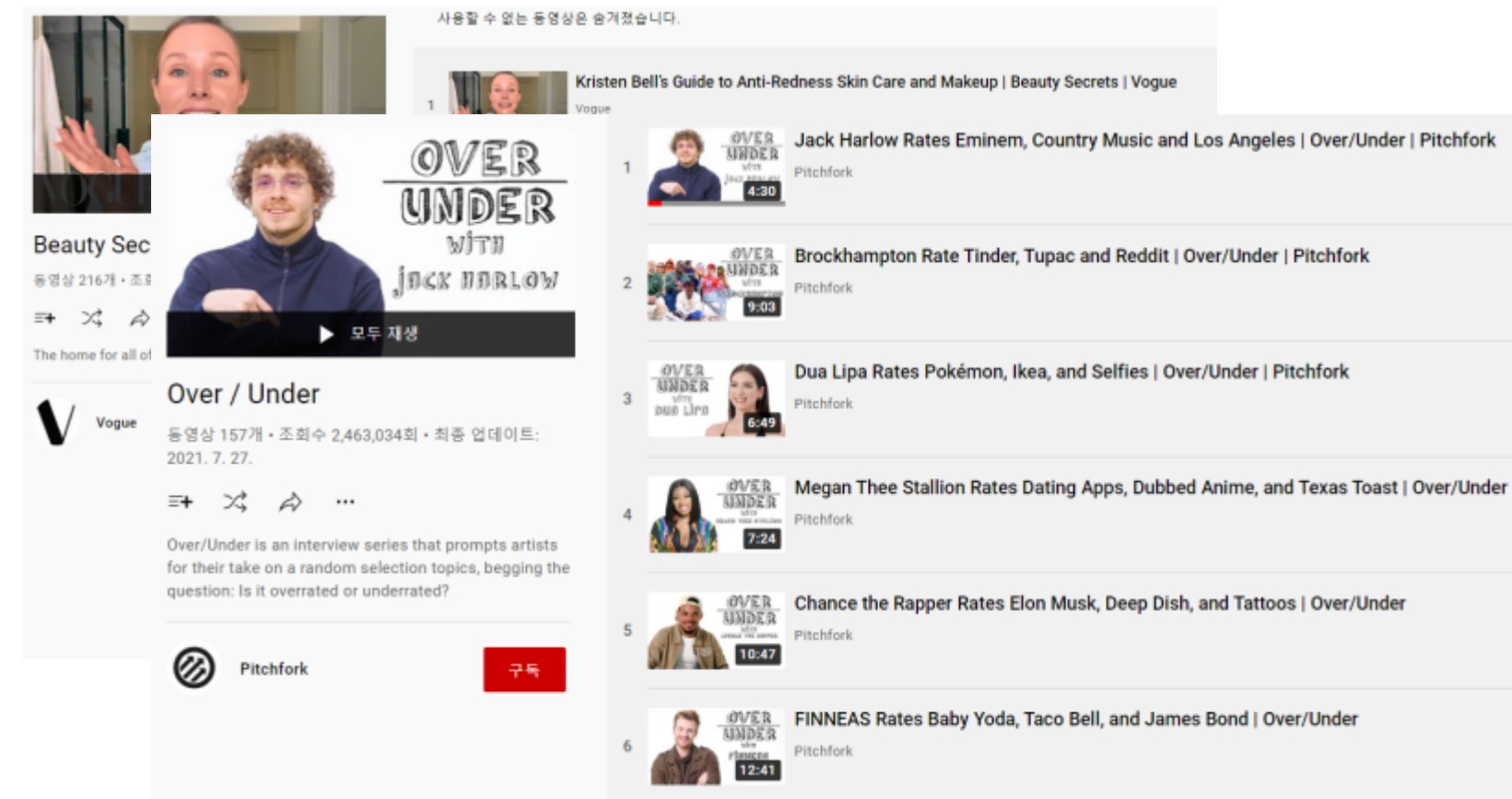
03

실제 적용 및 가능성

커스텀 데이터셋
한계점 및 활용방안

커스텀 데이터셋

부족한 시간 및 데이터로 인해, 영어로 학습되어 있는 기존의 모델을 유튜브 영상에 적용하기로 함



인터뷰, 리뷰 영상 등에 대해서 Lip Reading 모델을 적용

* Beauty Secrets (Vogue) : <https://www.youtube.com/playlist?list=PLztAHXmIMZFS9ZN7GTIZ2UOB2JmxiCd18>
 Over / Under (Pitchfork) : <https://www.youtube.com/playlist?list=PLrlyFA1NxiQZRuVIAPQbhTpyiO00vk8ww>

실제 학습시킨 결과

```
(wer=80.0) SHIT-I-DON'T-KNOW-MAN --> YOU-KNOW-WHAT
18/347 [>.....] - ETA: 1:59:34 - cer: 0.8533 - wer: 1.0265
(wer=140.0) HIS-FATHER-WAS-A-PITCHFORK --> AND-I-THINK-SHE'S-GOING-TO-DO
19/347 [>.....] - ETA: 1:59:10 - cer: 0.8529 - wer: 1.0462
(wer=100.0) IT'S-A-WILD-RIDE --> THANK-YOU-VERY
20/347 [>.....] - ETA: 1:58:44 - cer: 0.8509 - wer: 1.0439
(wer=100.0) I-SHOT-UP-THE-MEASLES-VACCINE-THE-OTHER-DAY --> IT-DOESN'T-SEEM-THAT
21/347 [>.....] - ETA: 1:58:20 - cer: 0.8458 - wer: 1.0418
(wer=100.0) I-SEND-HIM-A-TWEET-LIKE-SO-HE-COULD --> SIX-MONTHS-AND
22/347 [>.....] - ETA: 1:57:56 - cer: 0.8437 - wer: 1.0399
(wer=100.0) BECAUSE-HE-DOES-THAT-THING-WHERE-YOUR-MISSES --> AND-THIS
23/347 [>.....] - ETA: 1:57:33 - cer: 0.8436 - wer: 1.0381
(wer=100.0) THERE'S-NOTHING-BETTER-THERE'S-NOTHING --> DOESN'T-MEAN-THAT-THIS
24/347 [=>.....] - ETA: 1:57:10 - cer: 0.8359 - wer: 1.0365
(wer=100.0) IT'S-KIND-OF-RUBRA-JANA-I-CAN-IT'S-GOT-A --> I-DON'T-KNOW-WHAT-WE'RE-DOING-THIS-THING
25/347 [=>.....] - ETA: 1:56:48 - cer: 0.8351 - wer: 1.0351
(wer=100.0) KEEP-YOU-ON-THE-ROAD-FOR-A-YEAR-AND --> I'M-GOING-TO-FIND-OUT
26/347 [=>.....] - ETA: 1:56:25 - cer: 0.8326 - wer: 1.0337
(wer=88.9) I-MEAN-I-CAN-MAKE-THE-MOTIONS-OF-SWIMMING --> THANK-YOU-VERY-MUCH-FOR-THE-WOMEN
27/347 [=>.....] - ETA: 1:56:02 - cer: 0.8298 - wer: 1.0284
(wer=100.0) ECTOMORPHIC-PERSON-GROWING-UP --> SORT-OF-COMMODITY-STOCK
28/347 [=>.....] - ETA: 1:55:37 - cer: 0.8309 - wer: 1.0274
(wer=75.0) AND-APPARENTLY-I-HAVE-VERY-HIGH-BONE-DENSITY --> ABOUT-IT-I-HAVE-GOT-AN-IMPORTANT
29/347 [=>.....] - ETA: 1:55:15 - cer: 0.8258 - wer: 1.0178
(wer=83.3) WHEN-I-GET-IN-THE-WATER --> THAT-IS-THE-WORST
30/347 [=>.....] - ETA: 1:54:52 - cer: 0.8171 - wer: 1.0116
(wer=87.5) EVEN-IN-THE-AEGEAN-SEA-WHICH-IS-SALTIER --> THAT-IS-THE-QUESTION-OF-THE-CONTINENT
31/347 [=>.....] - ETA: 1:54:30 - cer: 0.8131 - wer: 1.0072
(wer=81.8) I-WOULD-HAVE-TO-SAY-THAT-HE-WAS-OVERRATED-EVEN-THOUGH --> WHAT-HAPPENS-IS-THAT-IT-WAS-SO-QUICKLY-FORGOTTEN
32/347 [=>.....] - ETA: 1:54:08 - cer: 0.8083 - wer: 1.0013
(wer=166.7) HE'S-ACTUALLY-DEMONSTRATED-SOMETHING-VERY-VALUABLE --> THAT'S-THE-TEMPERATURE-IS-GOING-TO-TRAVEL-ACROSS-THE-WORLD
33/347 [=>.....] - ETA: 1:53:45 - cer: 0.8111 - wer: 1.0215
```

Visual Model+Transformer+Extra Language Model

100.4%

Word Error Rate가 100%를 넘는 등
모델이 전혀 예측해내고 있지 못함

학습한 데이터

학습 된 데이터와 다른 언어 환경 (비속어 등)
실제 이용한 코드는 학습 및 parameter 값 변
경에 한계가 존재

한계점 및 활용방안

원래 계획과 많이 달라졌고, 좋은 결과를 얻지 못했지만 이러한 점들을 기대해볼 수 있다

01

추후 AIHub의 데이터가 공개될 시, 한국어 어휘를 학습시켜 문장 예측 모델 제작 가능

02

문장 처리 모델을 Transformer가 아닌 다른 모델 사용 시 Online 적용이 가능

03

이를 활용한 한국어 독순술 제공 어플리케이션 등의 가능성 존재

입술 데이터셋

- 입술 데이터셋은 국립국어원에서 배포한 '한국어 학습용 어휘'에서 자주 사용하는 1,000개를 선정하였으며, 200명의 배우를 섭외하여 단어를 발화하는 영상을 촬영하고 단어는 다양성을 고려하기 위해 11개의 품사에서 선정하였으며, 추가로 일상에서 자주 사용되는 생활용어를 추가하였다.

탄사	고유 명사	관형사	대명사	동사	보조 용언	부사	수사	의존 명사	형용사	명사	생활 용어	합계
12	21	25	25	151	2	61	35	28	69	514	57	1,000

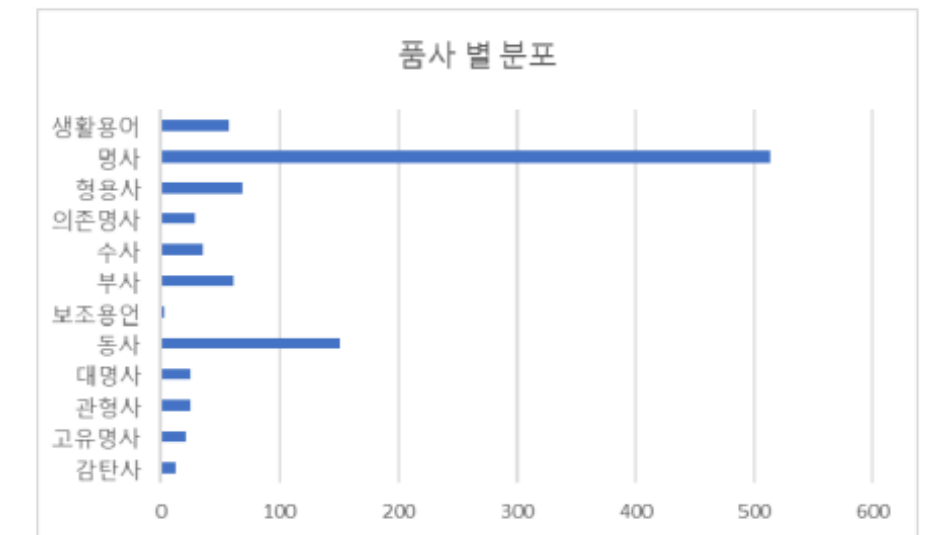


그림4 | 품사별 분류 분포

신체 말단 움직임 영상 데이터설명서

감사합니다.