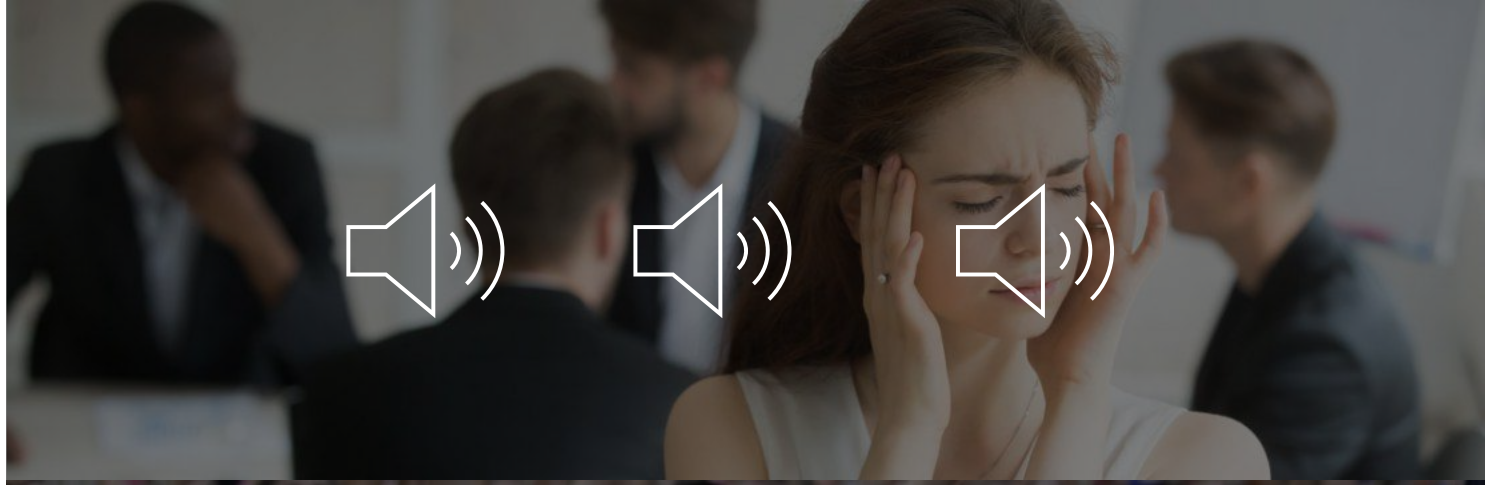
2021 데이터 청년 캠퍼스 [고려대]

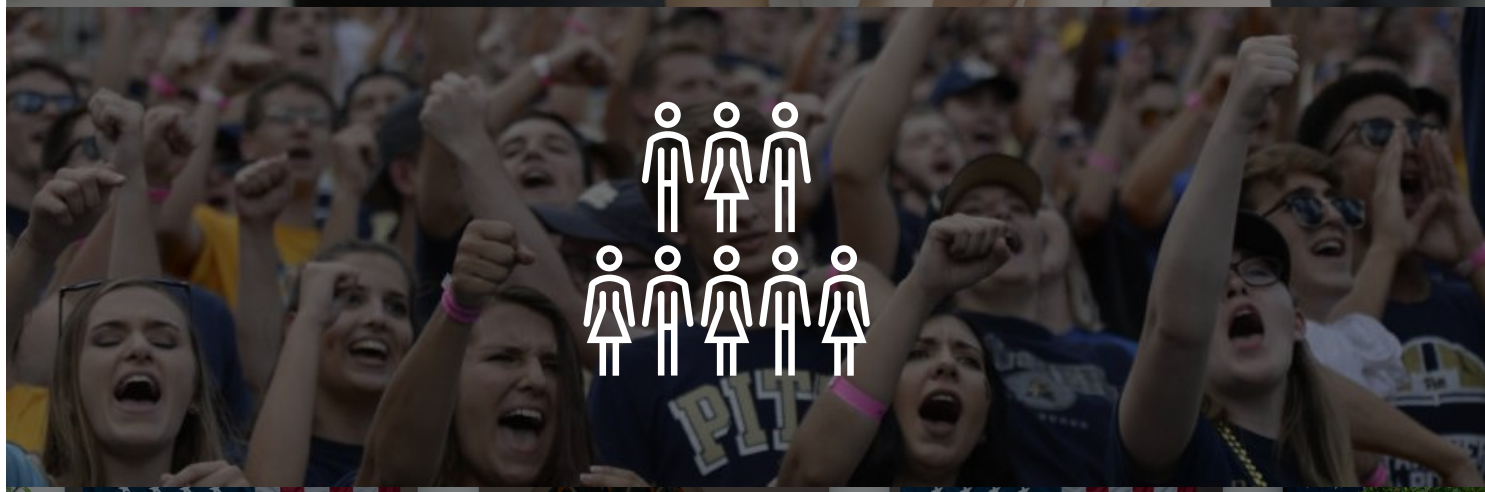
LIP 2 TEXT

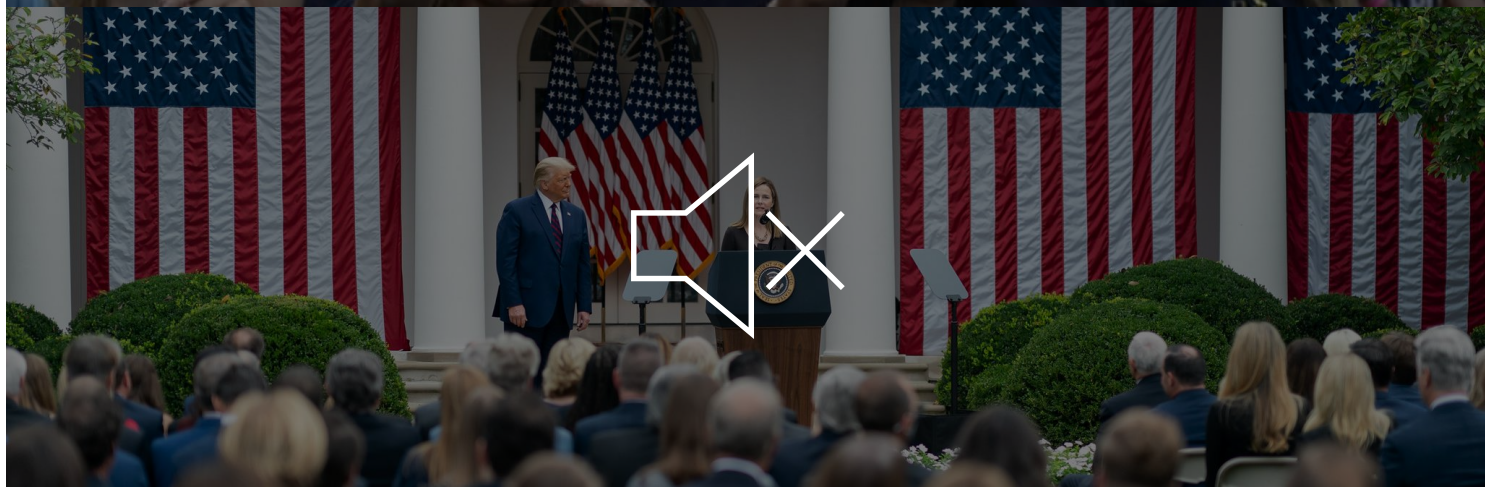
5조 이우준 안혜준 이태희 임준혁 송문영

Contents

주제 선정 배경







주제 선정 배경

주제 선정 배경

데이터 선정

뉴스 앵커 데이터

크롤링을 통한 데이터 셋 구축

장점

- **많은 양의 데이터 수급 가능**
  ↳약 30만개의 영상 데이터 접근 가능
- **문장형 영상데이터 수급 가능**
  ↳약 AI HUB 데이터와 달리 문장형 발음
- **아나운서의 정확한 입모양**
  ↳일반인보다 정확한 입모양 구현
  이를 통한 고순도 데이터 확보 가능
- **텍스트 데이터 제공**
  ↳뉴스 기사의 내용 텍스트로 제공
  이를 통해 차후 검증용 데이터로 쓸 수 있음

AI HUB
입 모양 데이터 셋

500,000개의 MP4 파일

장점

- **다양한 사람들 데이터**
  ↳뉴스 영상과 달리 다양한 사람들의
  발음 데이터 접근 가능
- **고해상도 영상데이터**
  ↳화자만을 녹화한 영상데이터로 영상
  잘라낼 시 뉴스 영상보다 고해상도 영상
  데이터 사용 가능

데이터 전처리

STEP 1

Sync Text and Video

학습의 Y 값 (TEXT)

학습의 X 값 (VIDEO)

FORCED ALLIGNMENT

↳Montreal Forced Aligner

* Only for 뉴스 데이터

STEP 2

Cut Video by Word

문장형으로 이루어진 영상

단어별로 영상을 자름

학습의 최소 단위 : 단어

* 모델 설명 파트에서 자세히 설명

* Only for 뉴스 데이터

STEP 3

Crop Lip from Video

입술 주변 턱 볼 하관 전체 이용

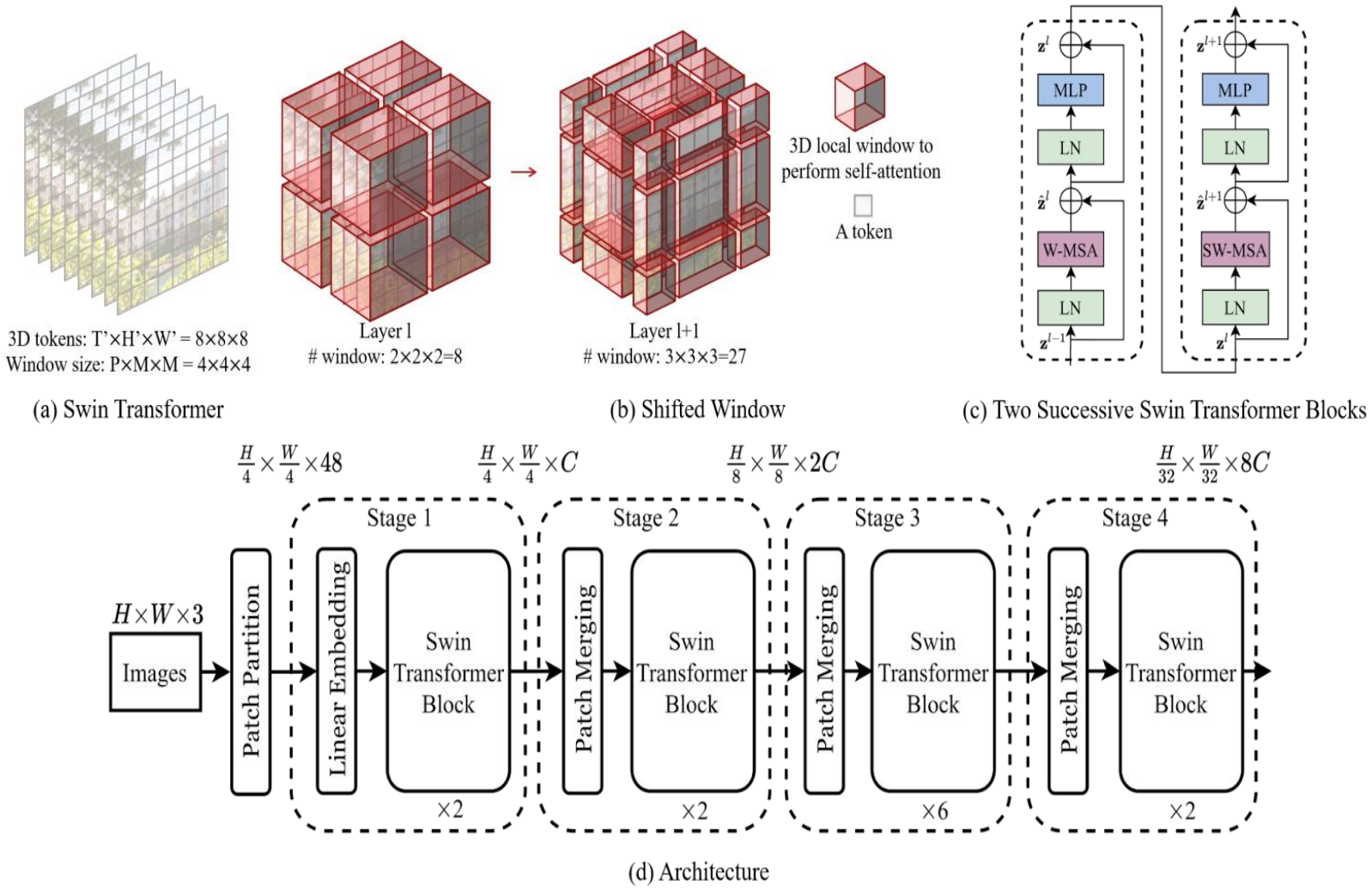
모델 설명

Vison Transformer

NLP

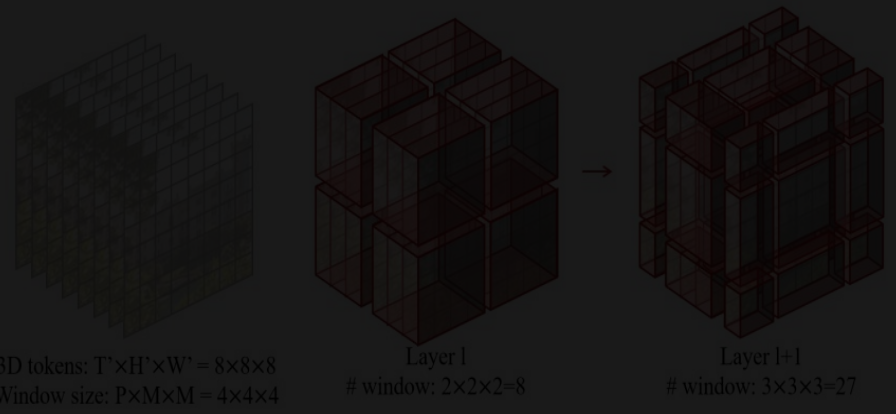
Vison Transformer

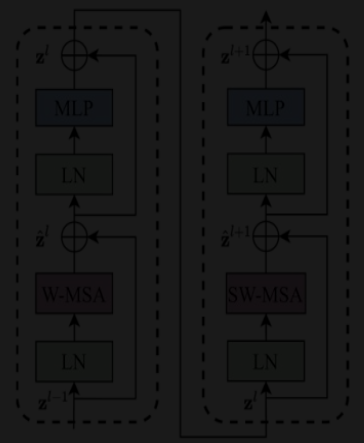
Video Swin Transformer



Vison Transformer

Video Swin Transformer





(c) Two Successive Swin Transformer Blocks

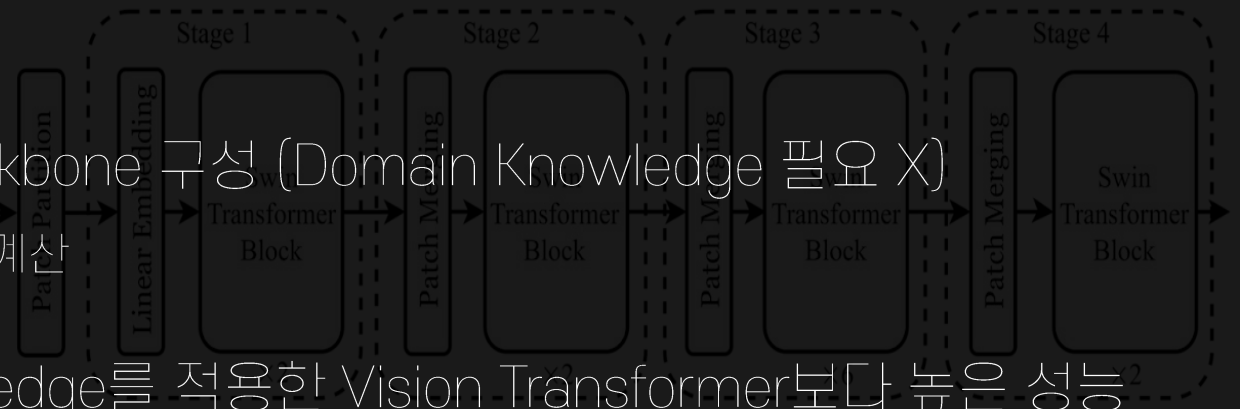
Input

단어에 따른 하관의 움직임(영상) (입 모양 + 턱 ,볼의 움직임)

Swin Transformer 이용

Patch Merging를 이용한 모델의 Backbone 구성 (Domain Knowledge 필요 X)

↳연속되는 프레임의 픽셀 간의 상관관계를 계산



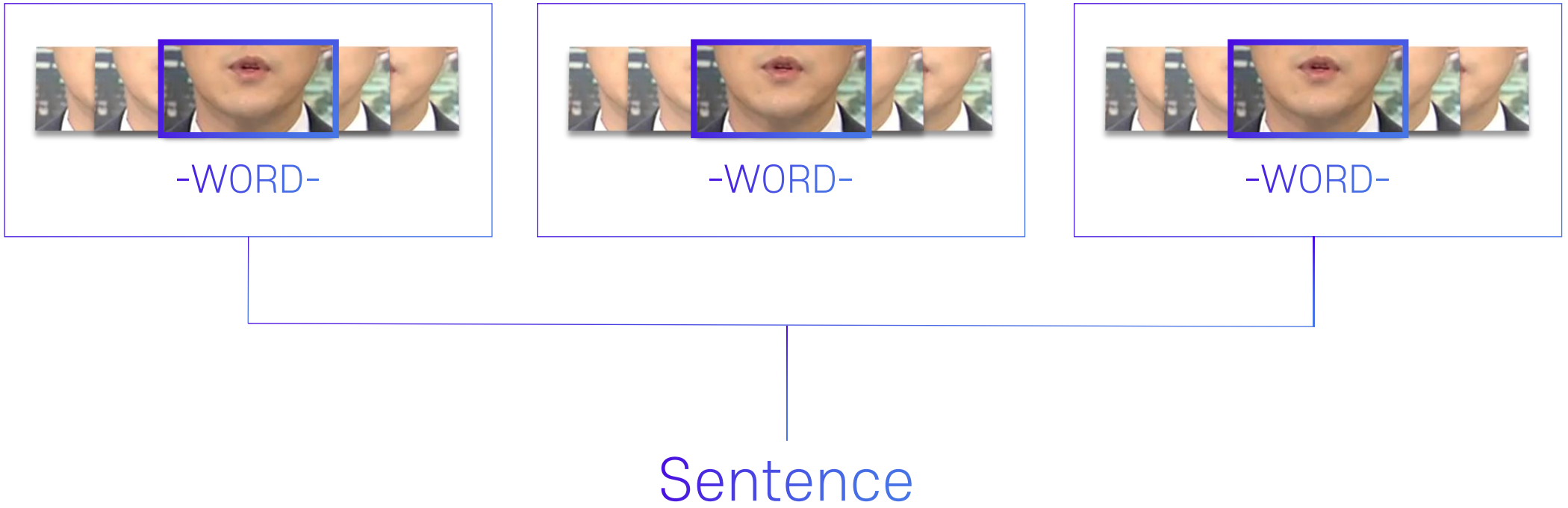
(d) Architecture

기존 CNN 기반 모델, Domain Knowledge를 적용한 Vision Transformer보다 높은 성능

↳ 행동 패턴 인지 부분에서 기존의 CNN, Vision Transformer 모델보다 높은 성능

Liu, Ze, et al. "Video Swin Transformer." arXiv preprint arXiv:2106.13230 (2021).

Natural Language Processing



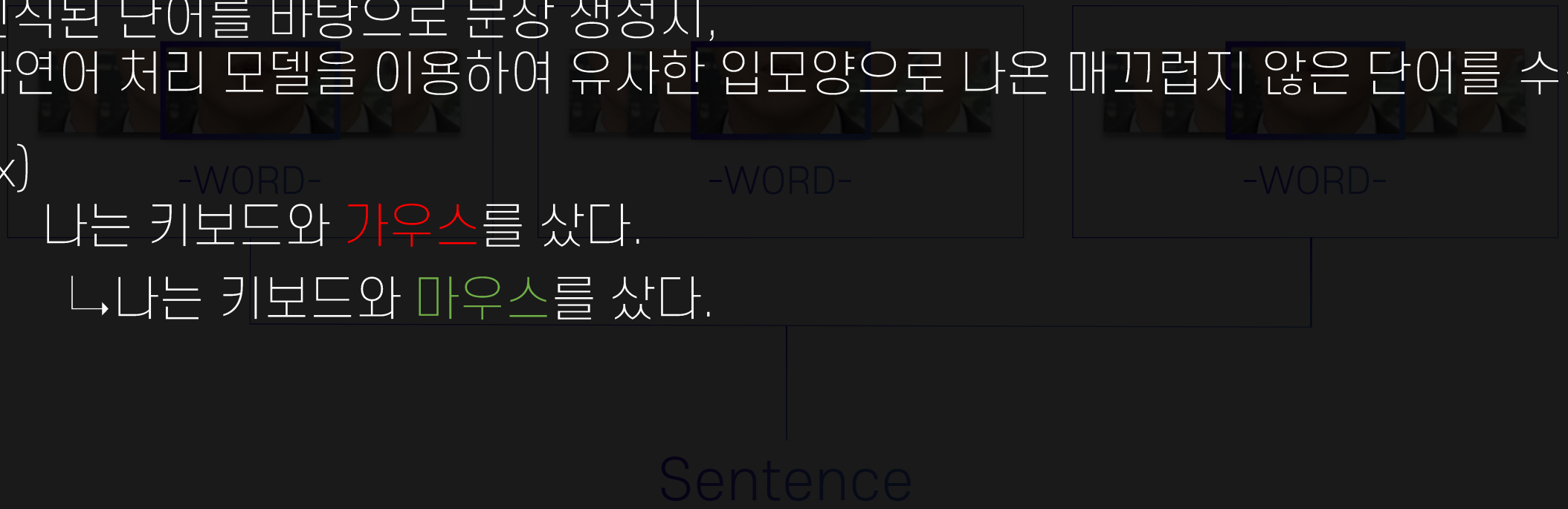
Natural Language Processing

인식된 단어를 바탕으로 문장 생성시,
자연어 처리 모델을 이용하여 유사한 입모양으로 나온 매끄럽지 않은 단어를 수정

Ex)

나는 키보드와 가우스를 샀다.

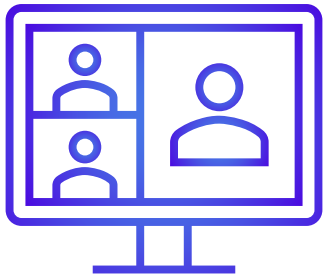
↳나는 키보드와 마우스를 샀다.



기대효과

청각장애인 독화법
소프트웨어
(스마트 글라스)

독화법 학습 소프트웨어



화상회의 프로그램
보조 소프트웨어

텍스트 + 감정전달

우주비행사
보조 통신수단

Q & A

Thank You