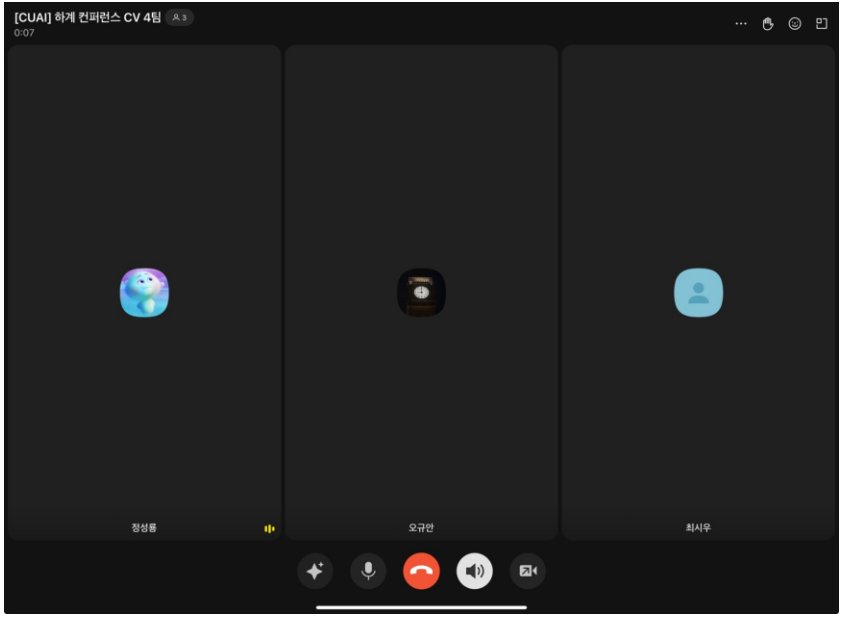
CUAI MM 프로젝트 2팀

2024.11.25

발표자 : 정성룡

스터디원 소개 및 만남 인증



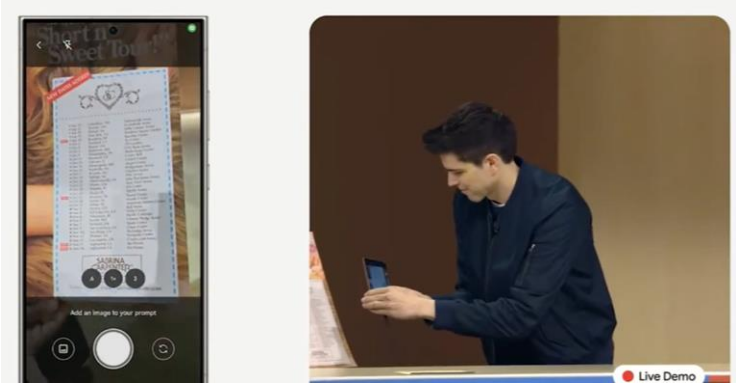
스터디원 1 : 오규안

스터디원 2 : 정성룡

스터디원 3 : 최시우

ChatGPT-4o, Gemini demo

ChatGPT-4o

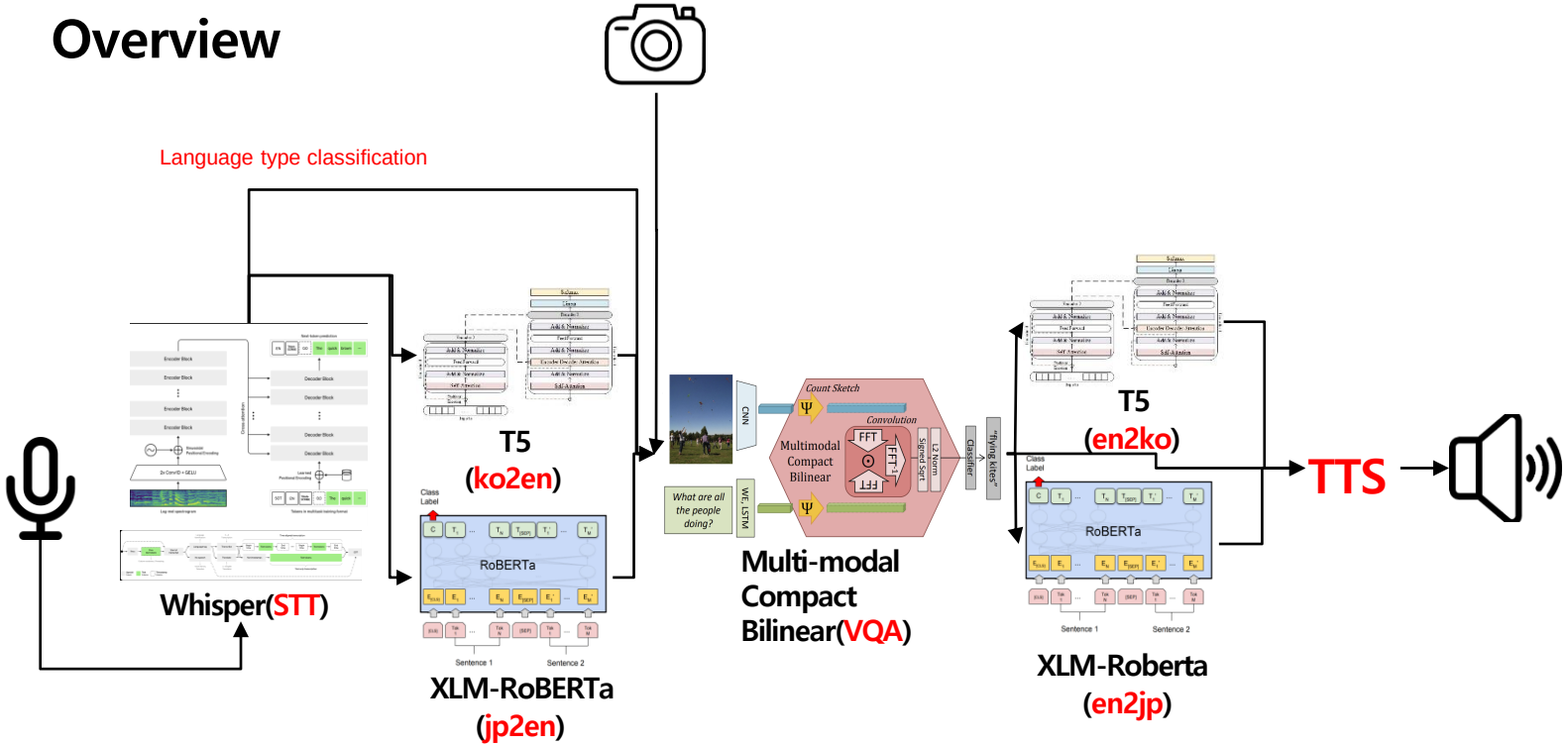


Gemini

Edge device 센서를 통한 실시간 VQA

>>> Edge device에서의 다국어 VQA 시스템

Overview

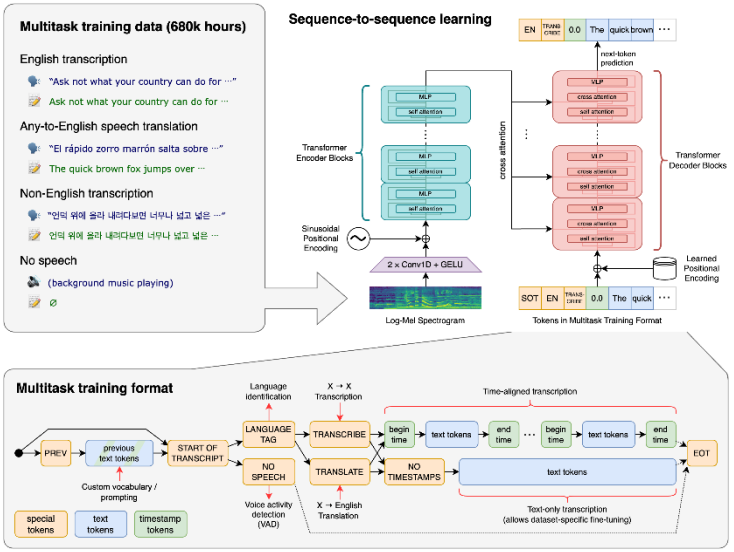


여러 모델을 조합하여 edge device에서의 실시간 vqa시스템 구축

Overview

VQA 데이터셋은 대부분 영어로 구성되어 있기 때문에, 한국어 또는 일본어가 입력으로 들어오면 이를 **영어로 번역한 후, 번역된 영어 텍스트를 VQA 모델에 입력**하여 텍스트를 생성하는 방식으로 구성

Whisper



STT에 사용하기 좋은 오픈소스 모델
(시간 처리에 적합한 경량화 필요)

빠른 whisper 모델:

- **Faster Whisper**
- **Distill Whisper**

- Distill Whisper는 영어 corpus만을 사용하여 한국어, 일본어 STT 작업이 불가능

- Faster Whisper 사용
- Hugging Face의 Distill Whisper 논문을 보며 한국어 및 일본어 corpus에 대한 distillation 시도 중

T5, XLM-Roberta

Machine Translation을 위한 모델

- 필요 요소:
  - Tokenizer에 source languag와 target language가 포함되어야 함
- 사용 모델:
  - XLM-RoBERTa (jp2en, en2jp)
  - KE-T5 (ko2en, en2ko)

T5, XLM-Roberta

translation dict
{ "en": "polar continent in the Southern Hemisphere", "ja": "地球の最も南にある大陸" }
{ "en": "general-purpose device for performing arithmetic or logical operations", "ja": "算術演算や論理演算を実行するための汎用デバイス" }
{ "en": "global system of connected computer networks based on IP addressing and routing protocols", "ja": "世界規模で接続されたコンピュータ・ネットワーク" }
{ "en": "song from the film and musical Mary Poppins", "ja": "映画「メリー・ポピンズ」の劇中で歌われる楽曲の名前" }
{ "en": "eleventh month in the Julian and Gregorian calendars", "ja": "グレゴリオ暦で年の第11の月" }

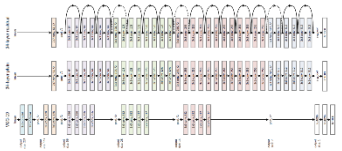
Skinner's reward is mostly eye-watering. string · lengths 10 706	스키너가 말한 보상은 대부분 눈으로 볼 수 있는 현물이다. string · lengths 4 220
Even some problems can be predicted.	심지어 어떤 문제가 발생할 건지도 어느 정도 예측이 가능하다.
Only God will exactly know why.	오직 하나님만이 그 이유를 제대로 알 수 있을 겁니다.
Businesses should not overlook China's dispute.	중국의 논쟁을 보며 간과해선 안 될 게 기업들의 고충이다.
Slow-beating songs often float over time.	박자가 느린 노래는 오랜 시간이 지나 뜨는 경우가 있다.

Fine-tuning을 위한 데이터셋

- 영어-일본어: 위키백과 문장 쌍
- 한국어-영어: 뉴스 문장 쌍

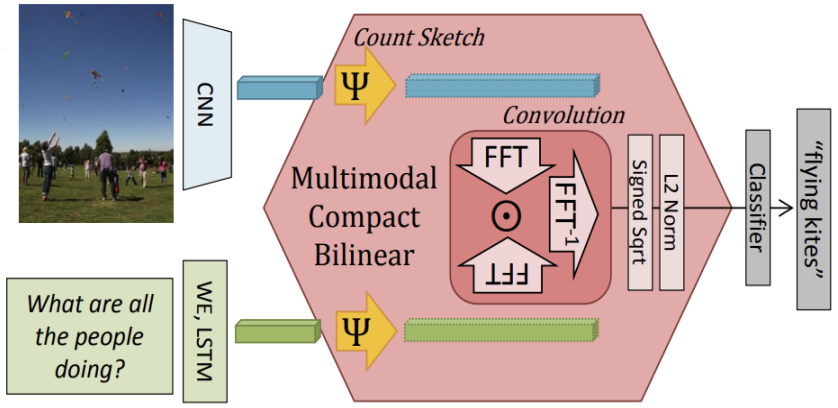
- 데이터셋 활용:
  - 같은 데이터셋에서 ko2en 및 en2ko(jp2en 및 en2jp)에 따라 source language와 target language 변경

Multi-modal Compact Bilinear



ResNet

TinyBERT



이미지 벡터 추출을 위해 ResNet 사용
텍스트 벡터 추출을 위해 TinyBERT 사용

Multi-modal Compact Bilinear

$$x_1 * x_2 = \mathcal{F}^{-1}(\mathcal{F}(x_1) \cdot \mathcal{F}(x_2))$$

각 모달리티에 대한 **Compact Bilinear Pooling** 적용
- 텍스트, 이미지 모달리티에 대해 독립적으로Compact Bilinear Pooling을 적용하여각 모달의 특징 벡터를 결합

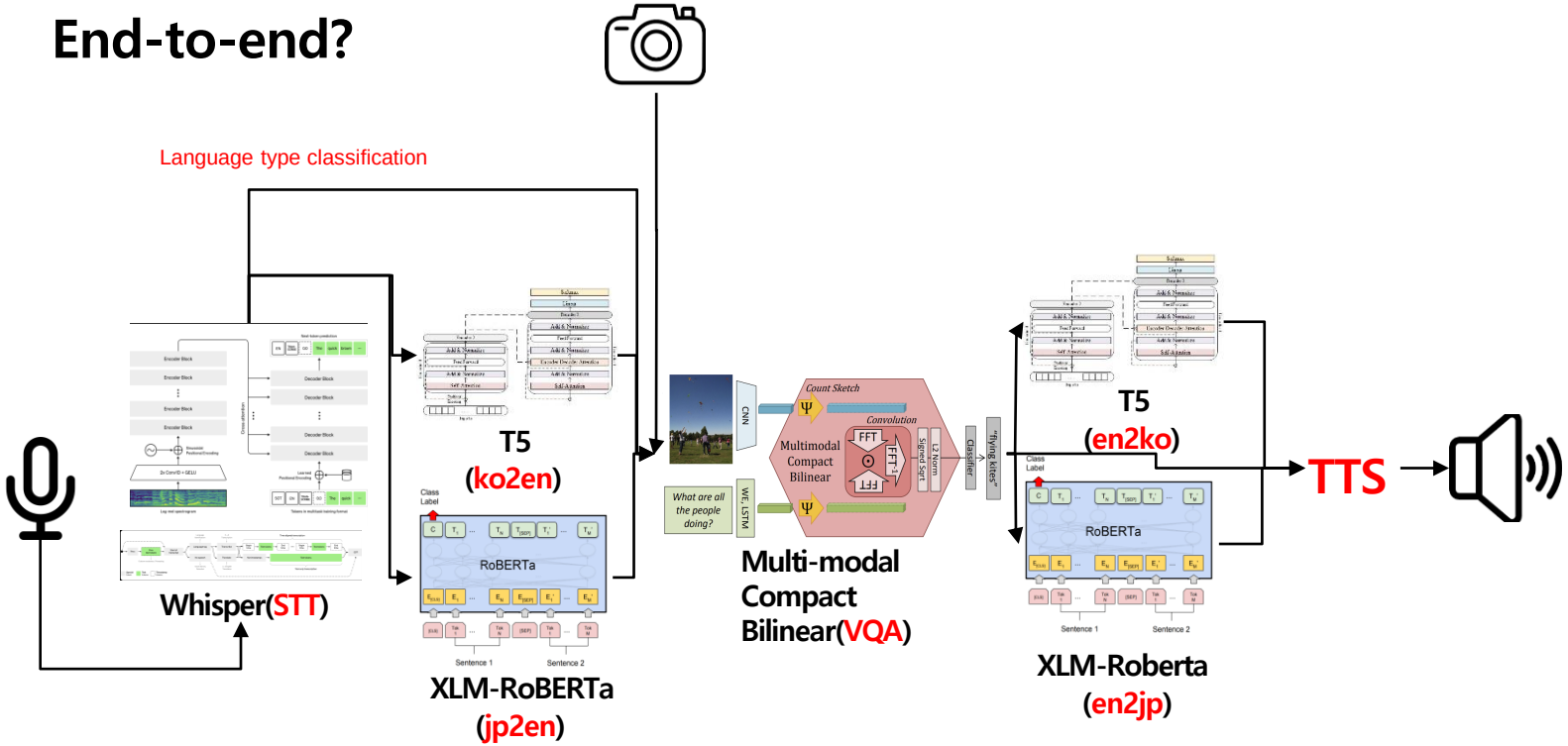
푸리에 변환을 통한 시간 복잡도 감소
- Bilinear Pooling에서 두 벡터의 내적을 계산할 때 발생하는 높은 시간 복잡도 문제를 해결하기 위해 푸리에 변환사용

Multi-modal Compact Bilinear

문제점
- Multi-modal Compact Bilinear (MCB) 적용 한 VQA 모델의 성능이 예상보다 저조
- 특히 센서를 통해 수집된 Raw Data에서 성능 저하가 심각

해결 방안
- VisualBERT, Flamingo 등 비교적 최신 모델을 활용예정

End-to-end?



현재 모델은 end-to-end 방식이 아니기 때문에, 각 작업 간에 오류가 전달되어 최종 성능에 부정적인 영향을 준다

감사합니다