Emotion-Guided Video Summarization

https://jeiyoon.github.io/spark/

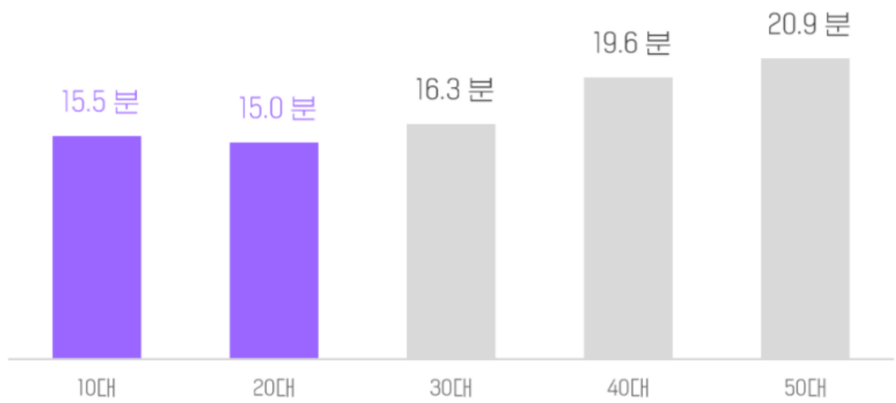
WAVE팀: 권기호, 이문기, 허주희, 박제윤
발표자: 박제윤

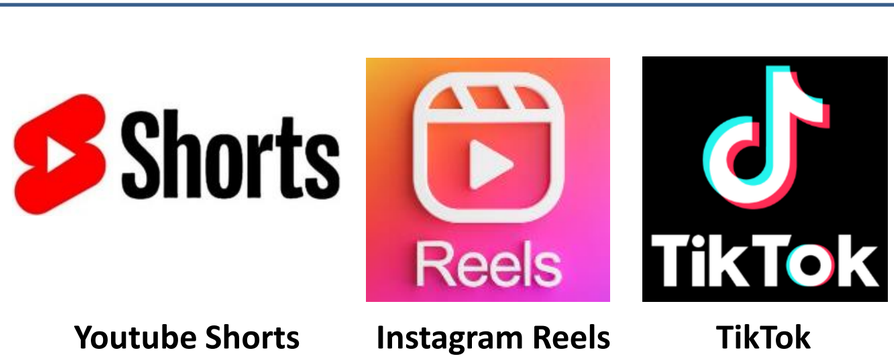
CONTENTS

01 | 아이디어 구상 및 제안 배경

1. 제안 배경

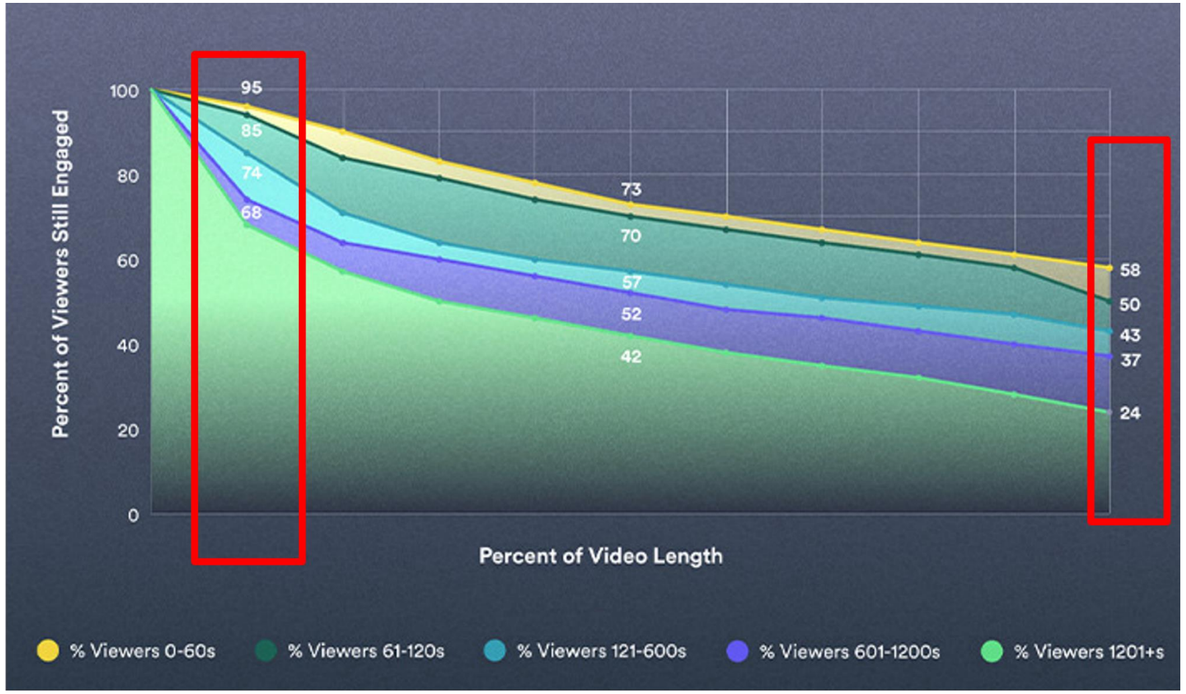
(1) 짧은 영상에 대한 수요 증가



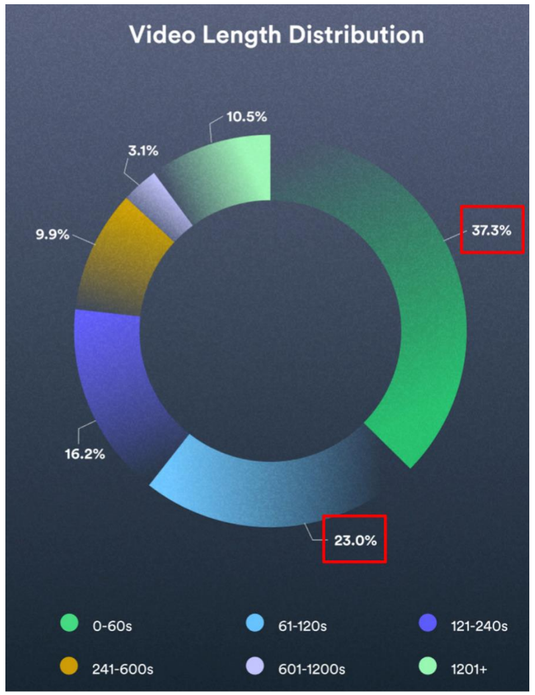


1. 제안 배경

(2) 비디오 길이에 관한 연구 (Vidyard, Video Benchmark Report 2021)



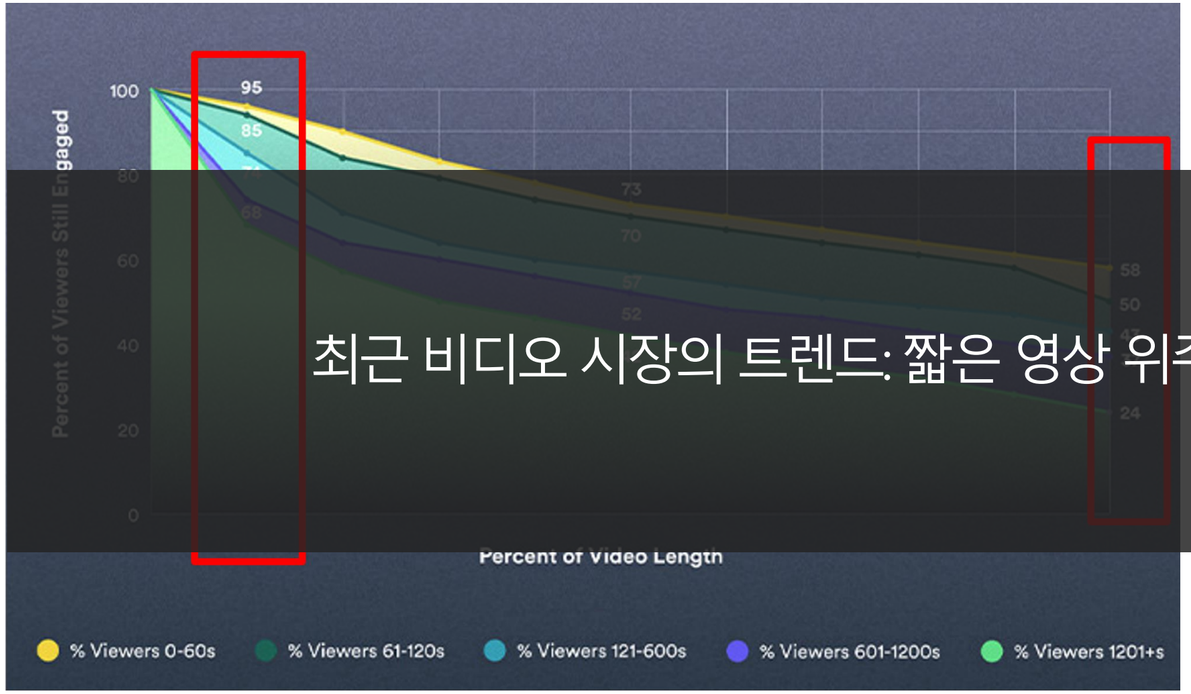
비디오의 길이에 따른 시청자의 평균 집중 정도 (출처: Vidyard)



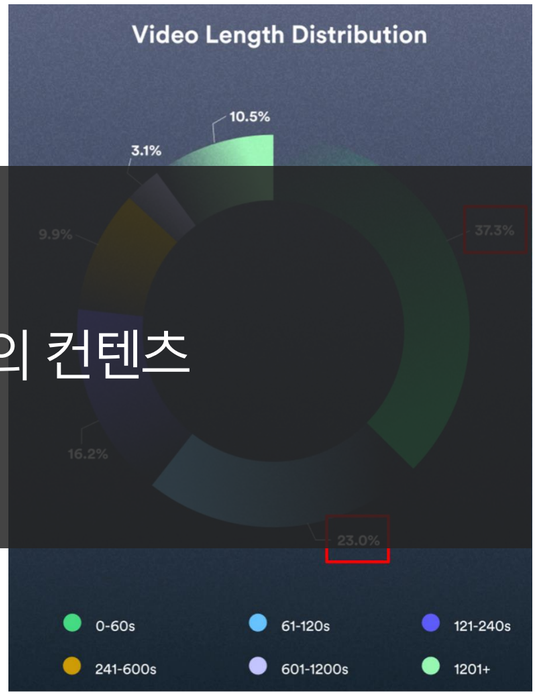
비디오 데이터의 길이 분포 (출처: Vidyard)

1. 제안 배경

(2) 비디오 길이에 관한 연구 (Vidyard, Video Benchmark Report 2021)



비디오의 길이에 따른 시청자의 평균 집중 정도 (출처: Vidyard)



비디오 데이터의 길이 분포 (출처: Vidyard)

최근 비디오 시장의 트렌드: 짧은 영상 위주의 컨텐츠

2 기존 모델들의 한계

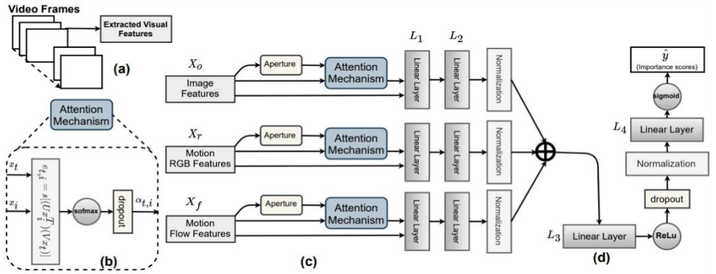
(1) Vision-Guided approaches vs. Language-Guided approaches

Vision-Guided Approaches

- 영상 내 visual feature들을 활용하여 중요한 장면 도출

- 영상의 움직임을 잘 포착하기 위해 (1) Optical flow 정보와 (2) RGB정보를 추출하여 활용

- 영상에서 중요한 캡션 정보와 오디오 정보를 활용하지 않고 시각 정보에만 의존하기 때문에 풍부한 요약이 어려움



Language-Guided Approaches

- 영상에서 자동으로 캡션정보를 추출하여 학습함

  e.g.) “레스토랑과 쇼핑센터가 포함된 장면들 보여줘”

- 사용자가 원하는 방향으로 영상을 추출할 수 있음

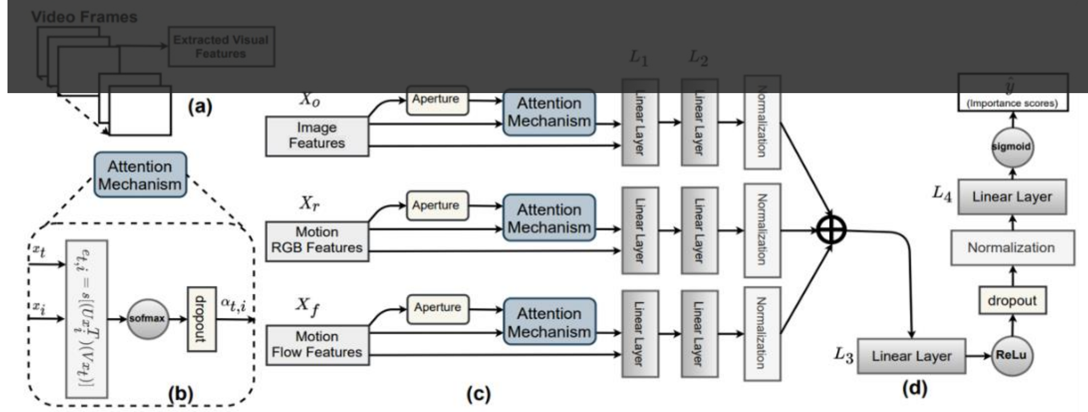
- 사용자 캡션이라는 주관적인 정보에 의존해야되기 때문에 어떤 영상인지 모르는경우 사용하기 어려움
- 언어정보는 음성정보에 비해 생동감이 떨어질 수 있음.
  e.g.) “천둥소리” vs “우르르쾅쾅”

2. 기존 모델들의 한계

(1) Vision-Guided approaches vs. Language-Guided approaches

Vision-Guided Approaches	Language-Guided Approaches
- 영상 내 visual feature들을 활용하여 중요한 장면 도출 - 영상의 움직임을 잘 포착하기 위해 (1) Optical flow 정보와 (2) RGB정보를 추출하여 활용	- 영상에서 자동으로 캡션정보를 추출하여 학습함 e.g.) “레스토랑과 쇼핑센터가 포함된 장면들 보여줘” - 사용자가 원하는 방향으로 영상을 추출할 수 있음

시각적인 정보와 주관적인 정보에만 의존하지 않고 생동감 있는 요약을 만들 수는 없을까?

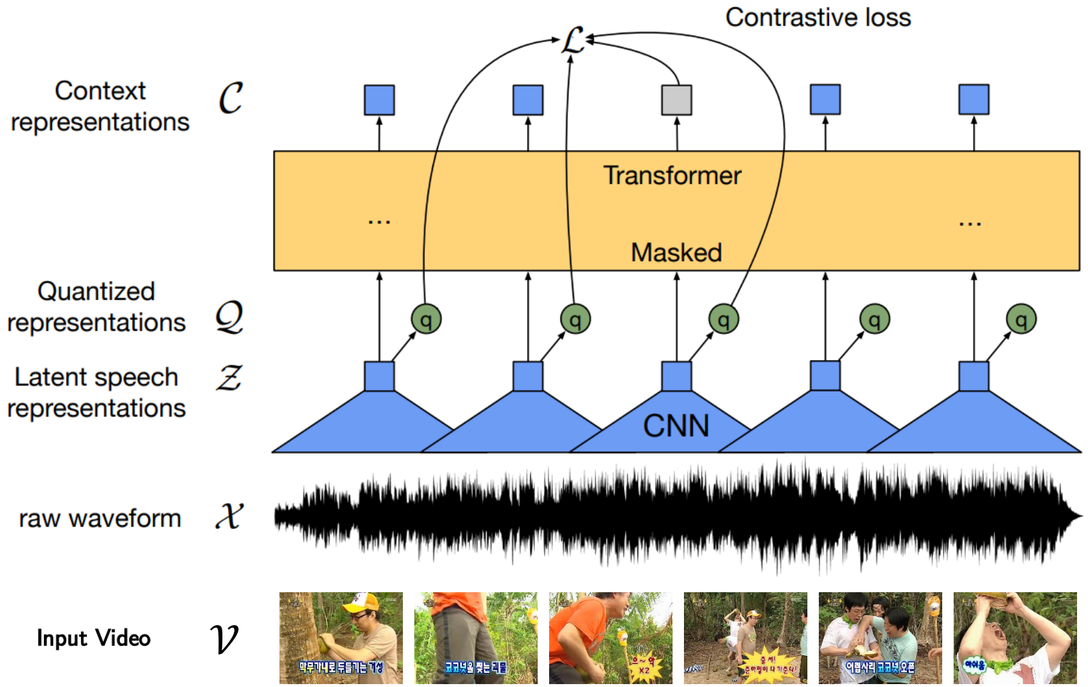


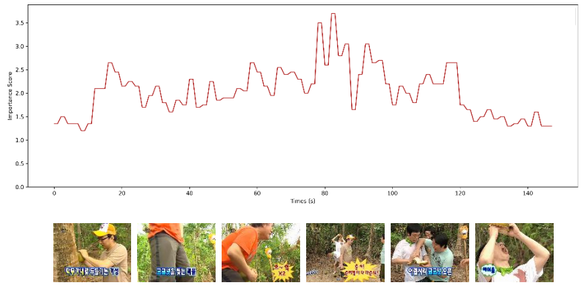
3. 가설 및 검증

(1) 가설: 영상에서 감정 정보가 몰리는 부분이 하이라이트일 것 이다.

3. 가설 및 검증

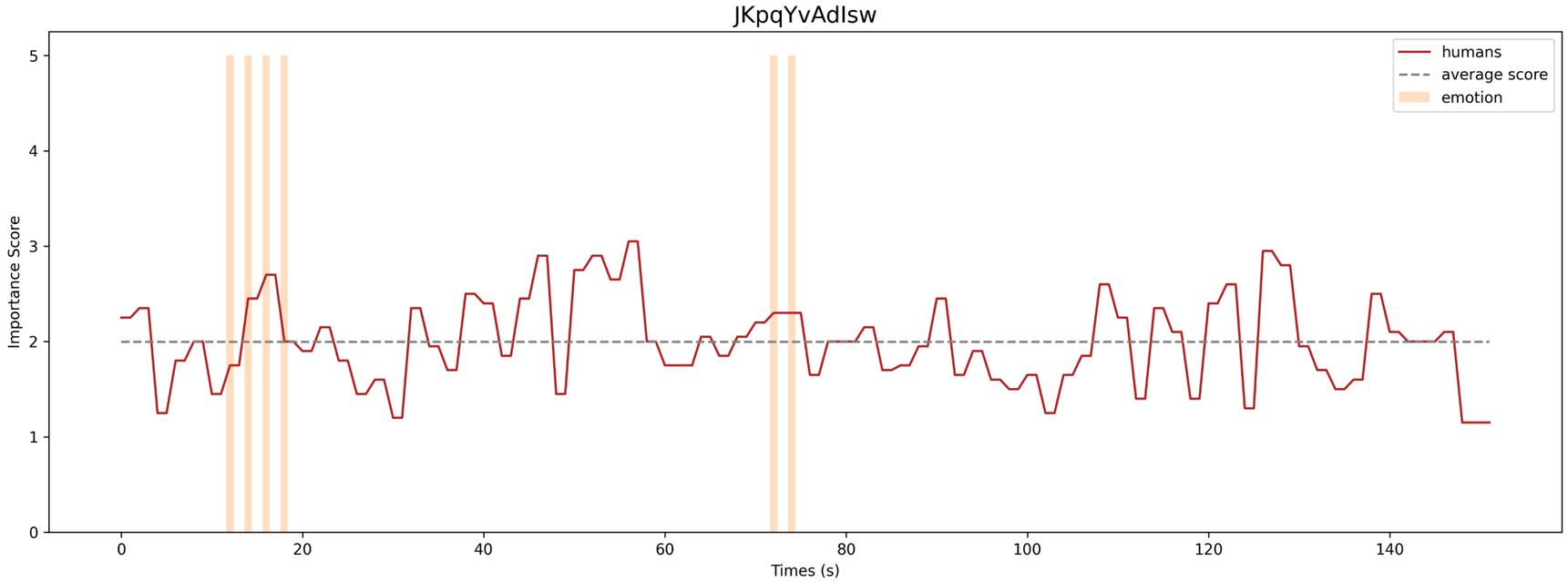
(2) 가설 검증:
- 1단계: Wav2Vec2 모델로 영상 (TVSum)의 음성에서 감정정보를 추출 하였음
- 2단계: 사람들이 실제 영상을 보고 매긴 하이라이트 점수 분포와 비교하였음





3. 가설 및 검증

(3) 결과: 50개의 영상을 분석. 대체로 점수가 높은 부분에 감정정보가 분포
- humans: 20명이 영상을 보며 시간에 따른 중요도를 입력한 것의 평균
- average score: humans의 평균. 전체 영상의 평균적인 점수



01 아이디어 구상 및 제안 배경

1. 제안 배경

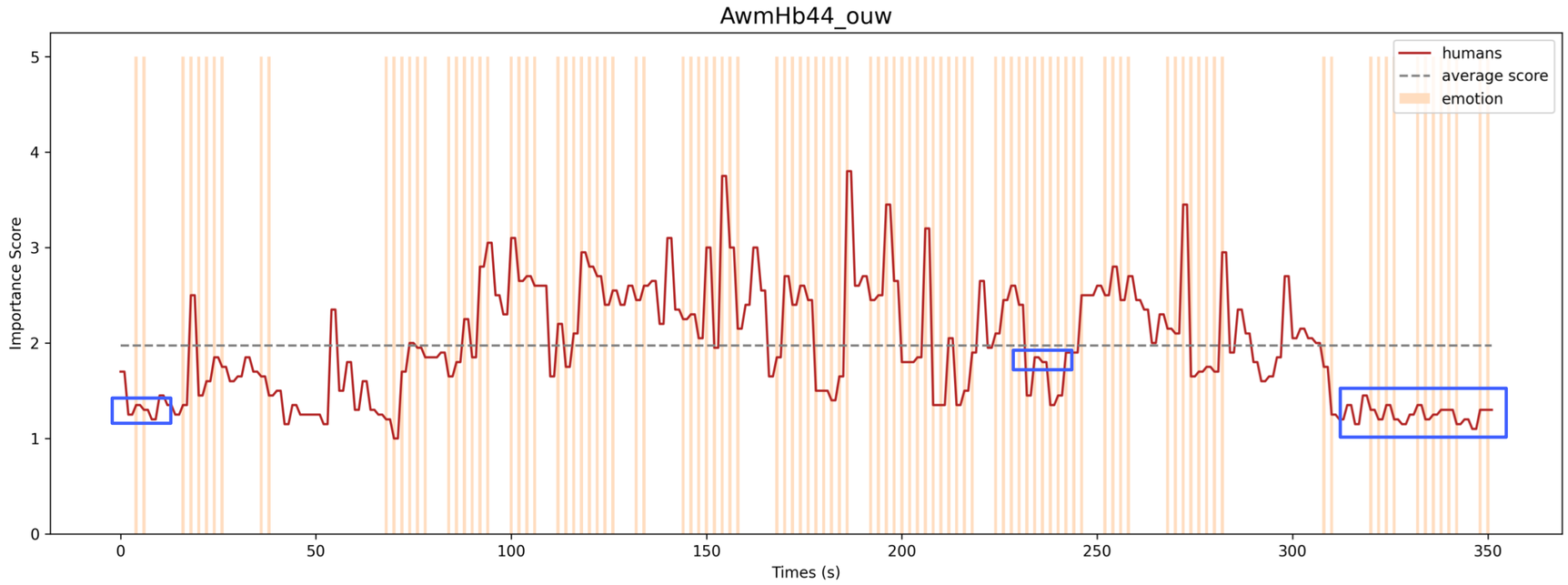
2. 기존 모델들의 한계

3. 가설 및 검증

4. 사용 데이터

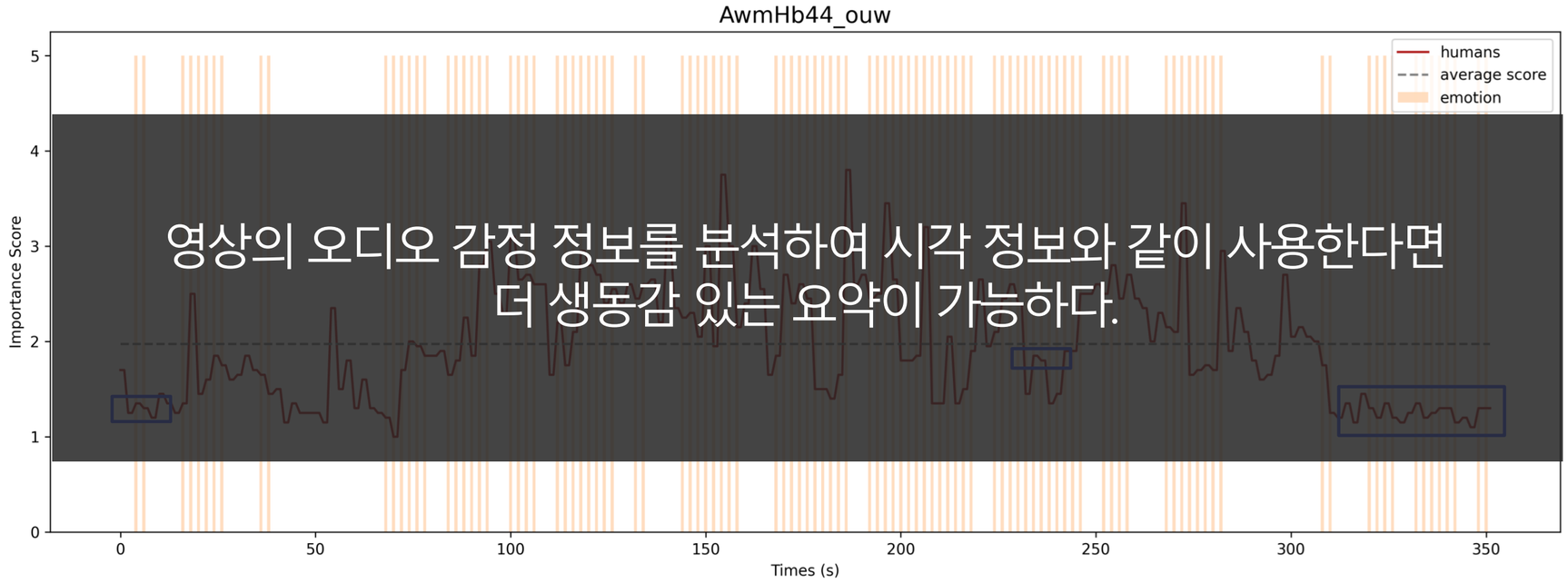
3. 가설 및 검증

(3) 실험 결과: 50개의 영상을 분석. 대체로 점수가 높은 부분에 감정정보가 분포함
- 아래 그래프에서는 48:34로 평균점수 이상에서 감정정보가 더 많이 몰려 있음
- 평균점수 이하인 부분들도 주변 점수보다 높은 부분들을 자주볼 수 있음



3. 가설 및 검증

(3) 실험 결과: 50개의 영상을 분석. 대체로 점수가 높은 부분에 감정정보가 분포함
- 아래 그래프에서는 48:34로 평균점수 이상에서 감정정보가 더 많이 몰려 있음
- 평균점수 이하인 부분들도 주변 점수보다 높은 부분들을 자주볼 수 있음



영상의 오디오 감정 정보를 분석하여 시각 정보와 같이 사용한다면
더 생동감 있는 요약이 가능하다.

4. 사용 데이터

(1) 출연연 데이터 목록

- [ETRI AI 나눔] 음성 감정인식 데이터셋
- [ETRI AI 나눔] 한국어 음성 감정 데이터셋 (KESDy18)

(2) 외부 데이터 목록

- [AI-Hub] 비디오 요약 영상
- SumMe (video)
- TVSum (video)
- RAVDESS (audio)
- IEMOCAP (audio)

4. 사용 데이터

(3) 출연연 데이터 구성
- ETRI AI 나눔의 데이터셋을 사용하였음
- 음성 감정인식 데이터셋, 한국어 음성 감정 데이터셋 (KESDy18)

음성 감정 인식 데이터셋

데이터 설명

- 압축 후 데이터 규모: 약 300.2MB

- 각 성우는 정적/ 동적 상태에서 네 가지 카테고리 감정 (중립, 행복, 슬픔, 분노)을 표현하며, 각 감정 카테고리 당 20문장의 한국어 문장을 발화

- 성우의 발화를 듣고 7가지 감정 레이블 (기쁨, 놀람, 분노, 중립, 혐오, 공포, 슬픔) 중 1개를 선택하고 각성도 (arousal)와 긍/부정도 (valence)를 평가하였음

한국어 음성 감정 데이터셋 (KESDy18)

데이터 설명

- 압축 후 데이터 규모: 약 9.22 GB

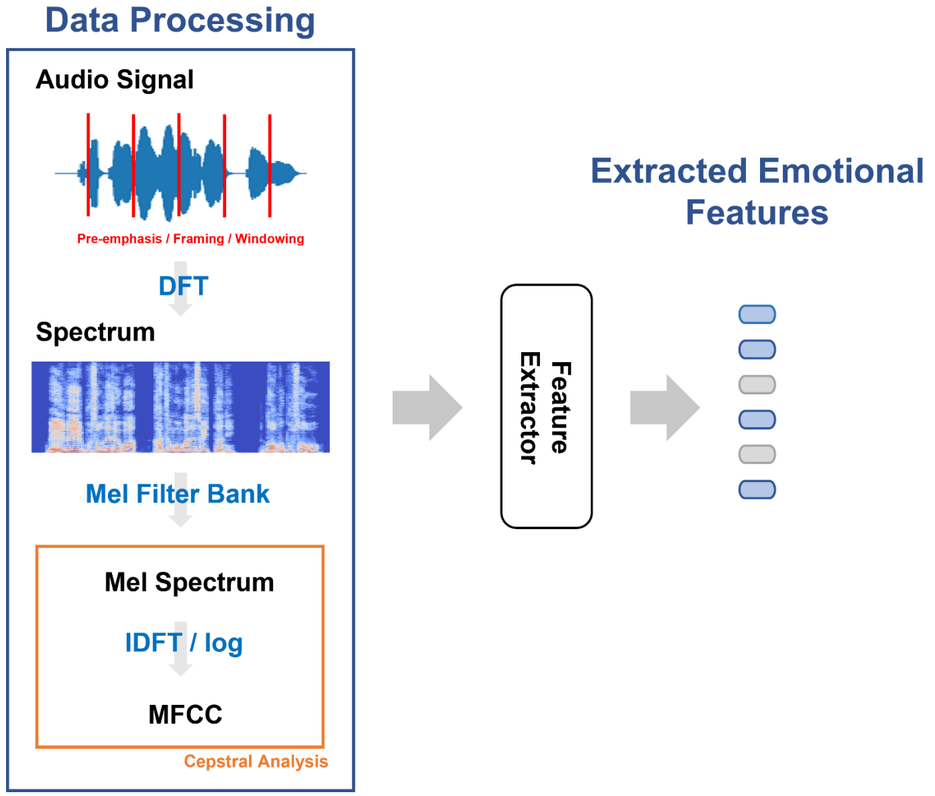
- 음성 약 10,000 문장, 코퍼스 27만 문장

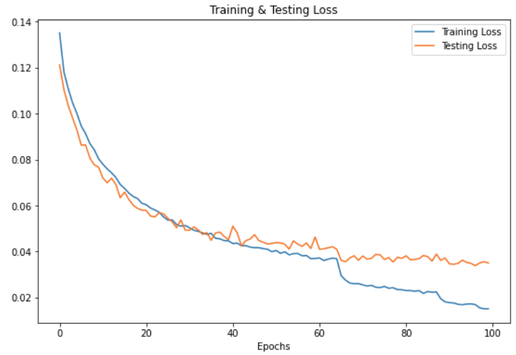
- 대표 감정 라벨:
분노,슬픔,불안,상처,당황,기쁨

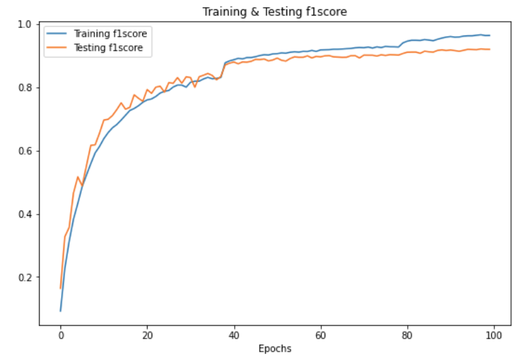
- 상세 감정 라벨: 좌절한, 짜증나는, 실망한, 후회되는, 두려운, 취약한, 배신당한, 외로운, 감사하는 등의 60여가지 감정

4. 사용 데이터

(4) AI 서비스 목적과 활용한 데이터 간 부합성
- 목적: 오디오의 감정정보를 활용하여 더 생동감 있는 영상 요약
- 영상의 오디오 감정 정보를 추출하기 위한 모델 학습에 사용







4. 사용 데이터

(5) 제공 데이터의 활용정도

- 영상 내 오디오 감정정보 추출을 위한 모델의 학습에 사용
→ 음성 감정인식 데이터셋, 한국어 음성 감정 데이터셋 (KESDy18)

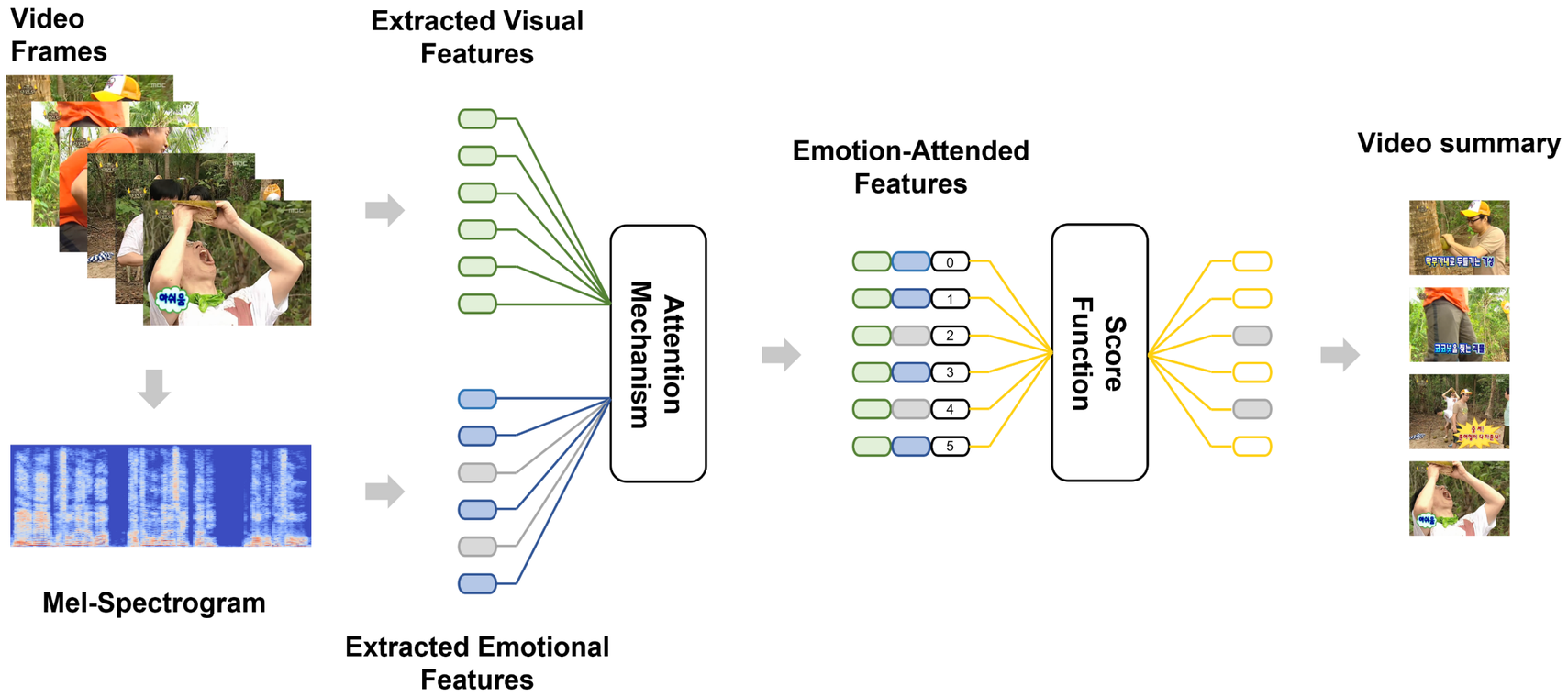
- 영상 내 시각정보 추출을 위한 모델의 학습에 사용
→ 비디오 요약 영상 데이터셋 (AI-Hub, 약 1TB)

- 영상에서 감정정보가 몰리는 부분이 하이라이트일 가능성이 높다라는 가설을 검증하기 위한 모델의 학습에 사용
→ SumMe, TVSum, RAVDESS, IEMOCAP

02 | 제안하는 AI 모델 및 서비스

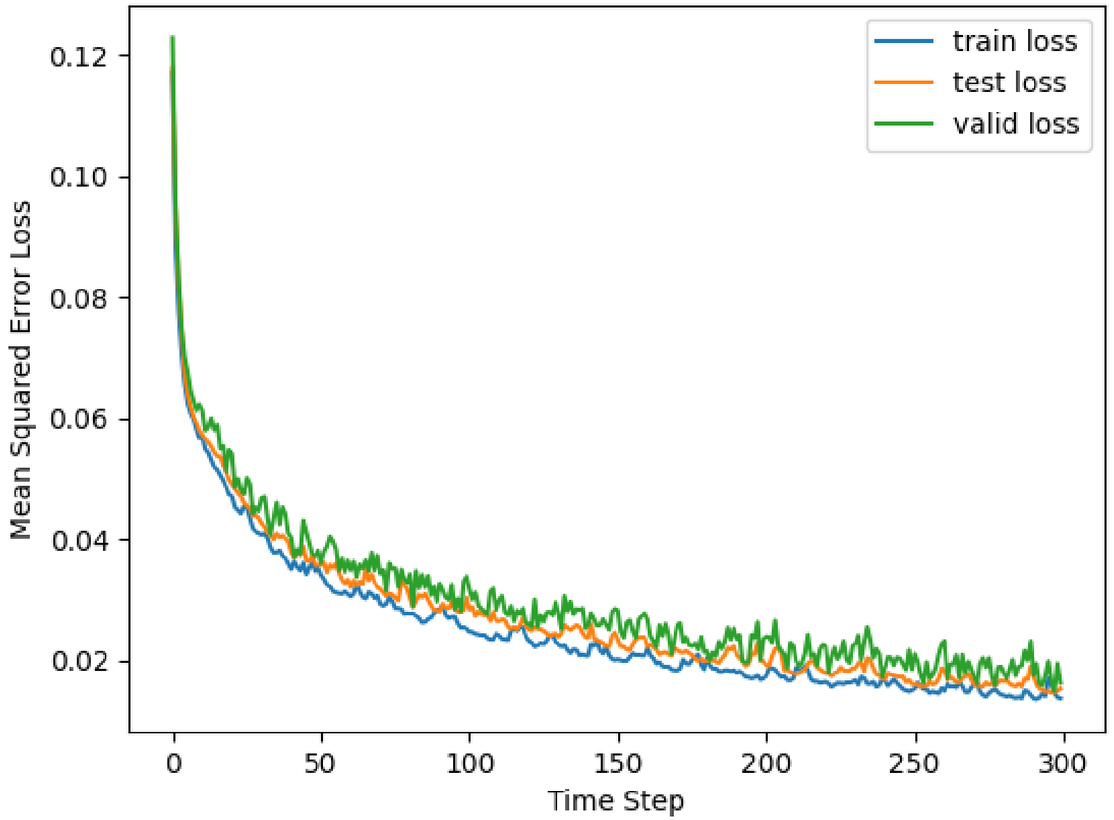
1. Emotion-Guided Video Summarization

- 모델 구조



2. 실험

(1) 학습 과정: Mean Squared Error Loss



2. 실험

(2) 실험1: 예측한 Importance Score와 실제 Human Annotation 을 비교함 (F1 score)

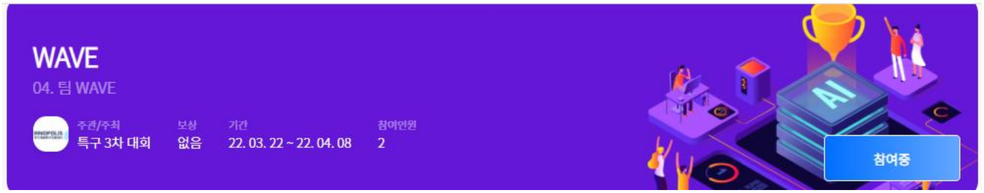
Method	Min.	Max.	Avg.
MSVA (Ghauri et al., 2021)	0.4169	0.5304	0.4527
Ours (EGVS)	**0.4884**	**0.6039**	**0.5273**

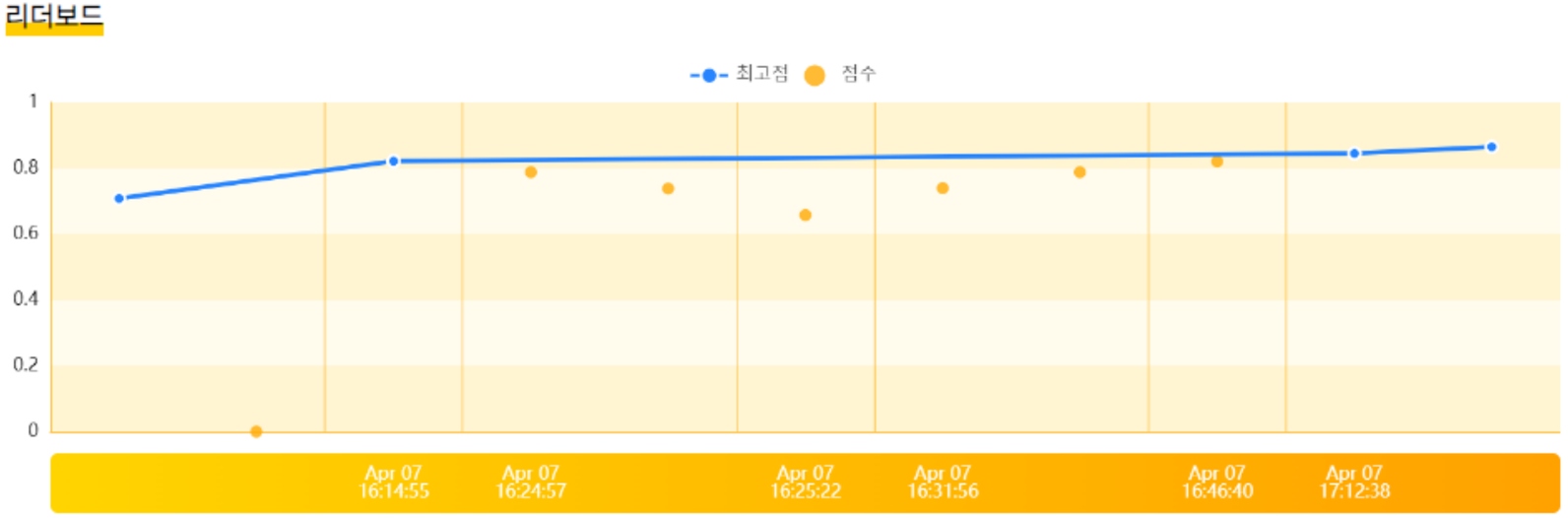
(3) 실험2: 전체 영상에서 점수가 높은 부분의 최대 15%를 하이라이트(1)로, 남은 부분을 하이라이트가 아닌부분(0)으로 레이블링한 후 맞추는 task를 진행함

Method	F1
MSVA (Ghauri et al., 2021)	0.7081
Ours (EGVS)	**0.8649**

2. 실험

(4) 리더보드
- 좋았던 점: SPARK를 진행하는동안 실시간으로 결과를 모니터링 하고 이전 결과들과 비교 분석할 수 있어서 성능 개선에 큰 도움이 되었음





2. 실험

(5) 체크 리스트
- 평가 메트릭 구현 [✓]
- 알고리즘 개발 [✓]
- 리더보드 제출 1회 이상 [✓]

(모든 코드는 공개하여 다른 사람들도 사용 가능함: https://Jeiyoon.github.io/spark/)

```python
##################################################
# F1 Score
##################################################
import os
import sys
import pandas as pd
import numpy as np
import math
#############################################################################
# 편집할 구간: 채점에 사용할 함수 정의
# evaluate_summary: 점수를 계산할 함수입니다. 정답(user_summary), 예측(machine_summary)을 인자로 입력받아 score를 반환합니다.
def evaluate_summary(machine_summary, user_summary, eval_metric='avg'):
    """Compare machine summary with user summary (keyshot-based).
    Args:
    --------------------------------
    machine_summary and user_summary should be binary vectors of ndarray type.
    eval_metric = {'avg', 'max'}
    'avg' averages results of comparing multiple human summaries.
    'max' takes the maximum (best) out of multiple comparisons.
    """
    machine_summary = machine_summary.astype(np.float32)
    user_summary = user_summary.astype(np.float32)
    n_users, n_frames = user_summary.shape
```

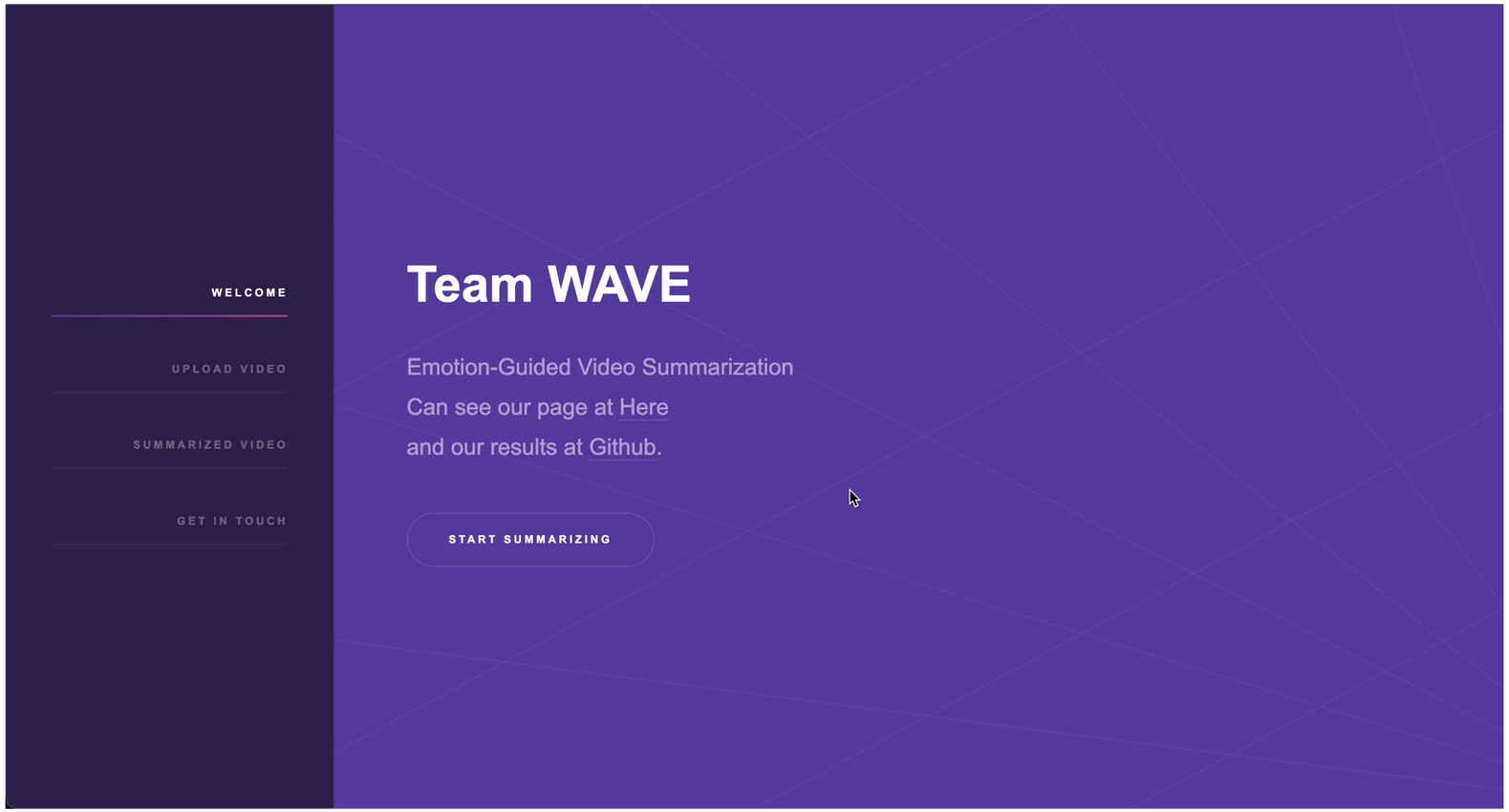
3. 요약 생성 및 비교

- 감정정보가 추가된 representation으로 학습할 경우 훨씬 더 감정표현이 많은 부분을 집중해서 요약해주기 때문에 훨씬 자연스럽고 재밌음

02 제안하는 AI 모델 및 서비스

4. 데모

- 임의의 영상을 넣으면 생동감있는 요약을 해주는 데모를 만들어 실제 서비스화를 염두해두었음



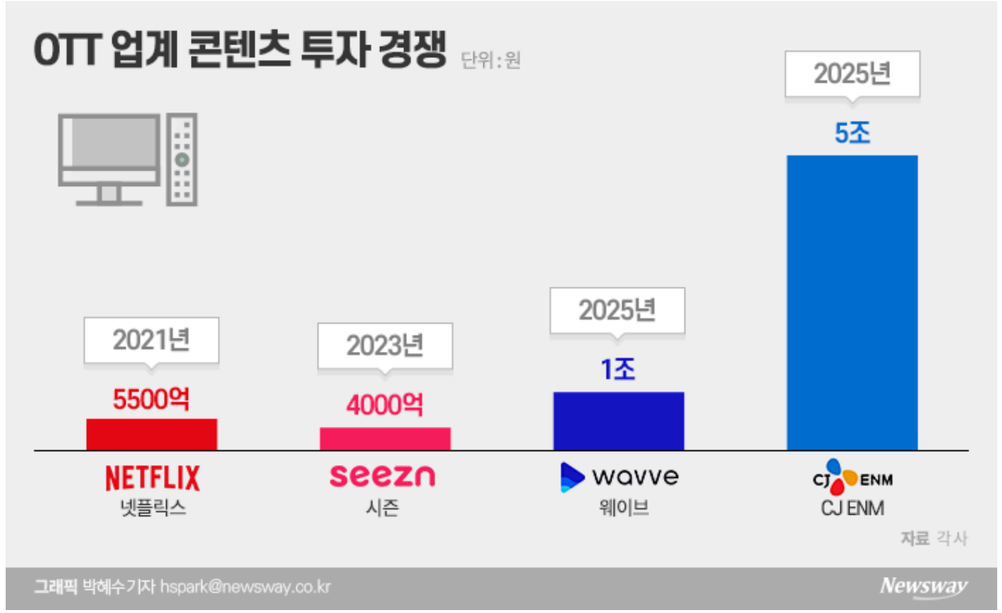
03

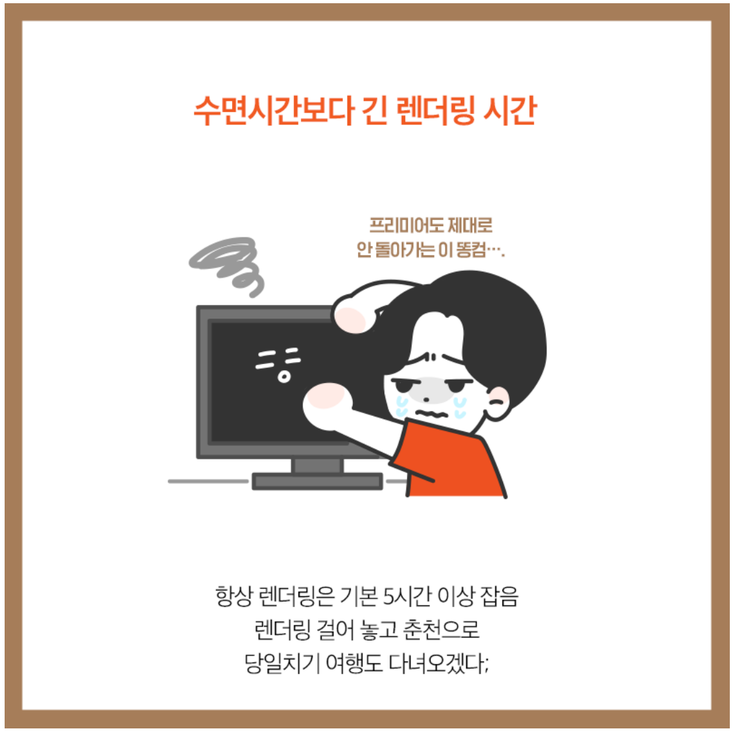
시장성 및 아이디어 현실성

1. 시장성

- 방송사 및 OTT 기업과의 연동을 통한 문화 컨텐츠 재창출에 기여

- 소규모/1인 영상제작자에 대한 업무 경감 지원





2. 현실성

- 가설 검증을 통한 입증, 기존 모델을 뛰어넘는 성능, 서비스를 위한 데모 제작

- 요약의 수요를 입증하는 YouTube 요약 영상의 조회수

안 본 사람은 있어도 한번 본 사람은 없다는.. 후유증이 심각한 드라마 ‼달의 연인 보보경심 려 한방에 보기

조회수 511만회 • 2개월 전

티비요정

달의연인보보경심려 #넷플릭스 #아이유 #이준기 #강하늘 #홍종현 #백현 #지수 #남주혁 #강한나 #진기주 #서현 #넷플릭스 #Netflix ...

한지민을 버리고 과거로 돌아가 부자 와이프를 선택하고 살다 다시 한지민이랑 불륜을 저지른 남자 [아는와이프] #즐거움앳홈파티

조회수 908만회 • 1년 전

고몽

고몽의 최애 드라마 리스트에 들어가 있는 아는와이프입니다. 진짜 한지민 짱입니다. 강한나도 짱입니다. #아는와이프 #한지민지성 ...

감사합니다!

권기호

이문기

허주희

박제윤

AI SPARK
CHALLENGE

권기호, 이문기, 허주희, 박제윤
발표자: 박제윤