BOSTON UNIVERSITY

YouTube

YouTube-8M Video Understanding Challenge / 5-SEC

Cagri Yoruk, Zeyu Song

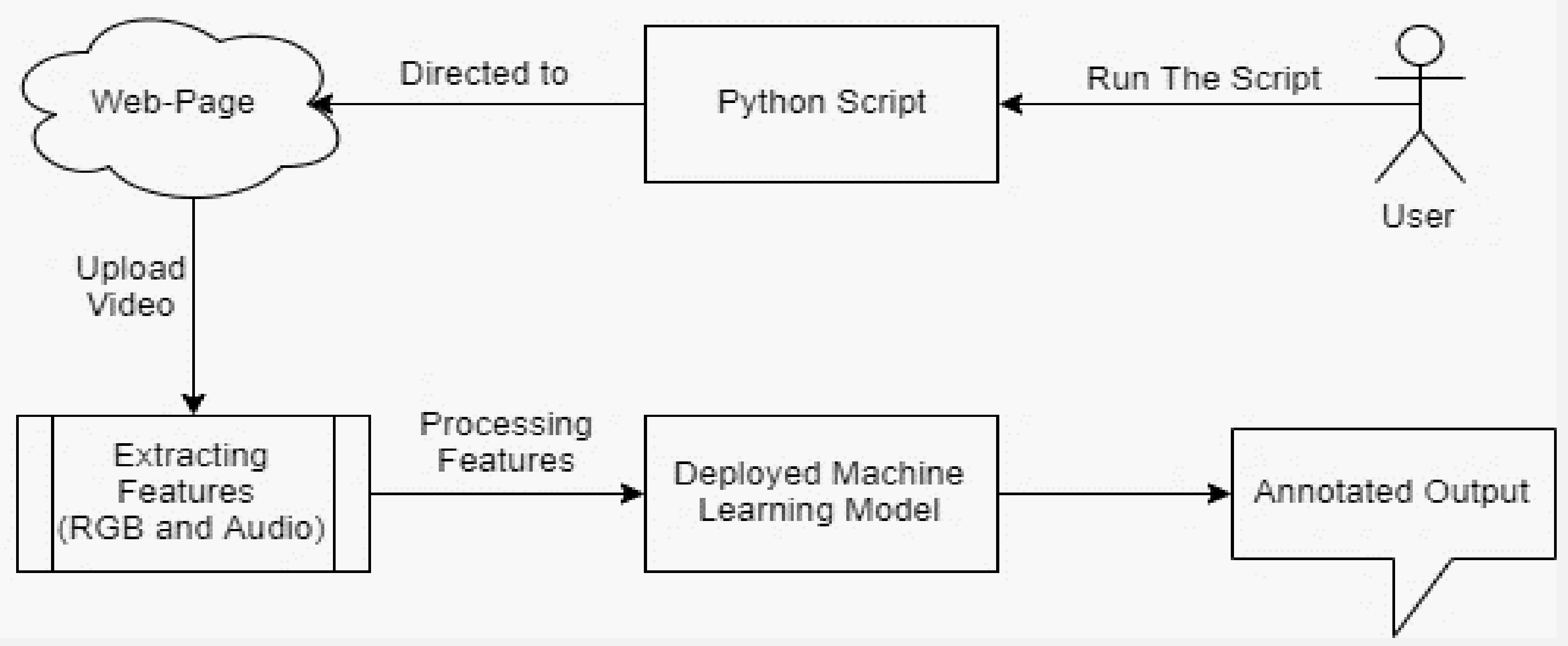
Why 5-Sec?

Our ultimate goal is to make people get 5 Second highlights of their videos.

Mission

Be able to search for the moment in any video without providing description or such metadata.

System Design



YouTube 8M Dataset

YouTube-8M Segments dataset is an extension of YouTube-8M dataset with human-verified segment annotations.

6.1 Million
Video IDs

350,000
Hours of Video

2.6 Billion
Audio/Visual Features

3862
Classes

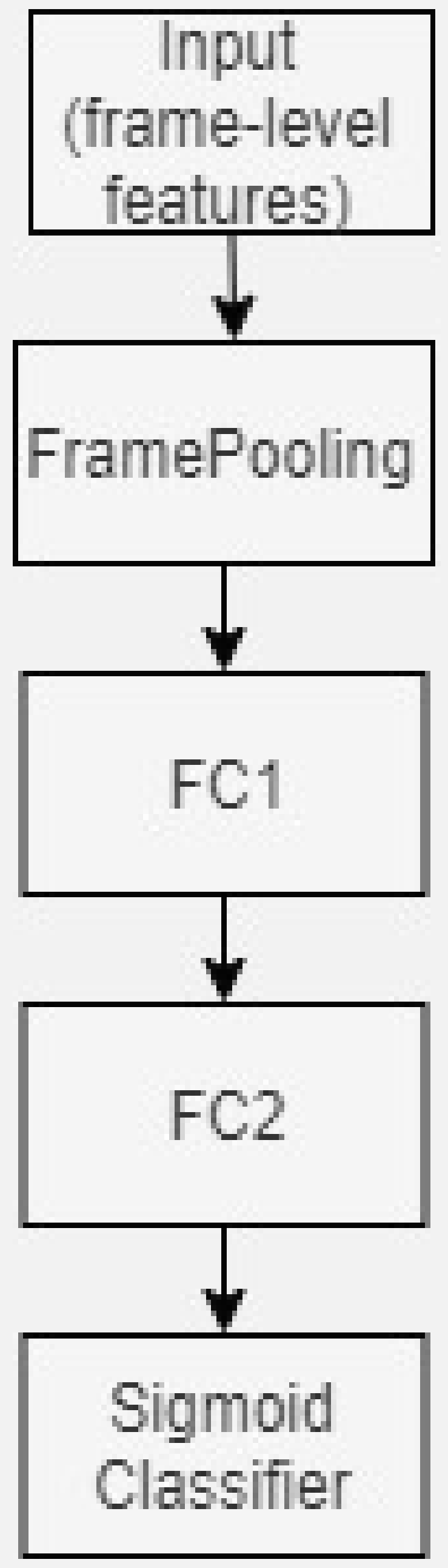
3.0
Avg. Labels / Video

237K
Human-verified Segment Labels

1000
Classes

5.0
Avg. Segments / Video

Deep Bag Of Frames



- For each sample in a training dataset there is a set of frame-level features and ground-truth video-level labels.
- Input data is sent to Frame Pooling layer, where pooling between time frames of each sample is applied.
- We use max-pooling to get one feature-vector from all time-based frame-level features of each input sample.
- After FramePooling layer two FC layers are used. And on the top level we use sigmoid classifier.

Sample Submission File

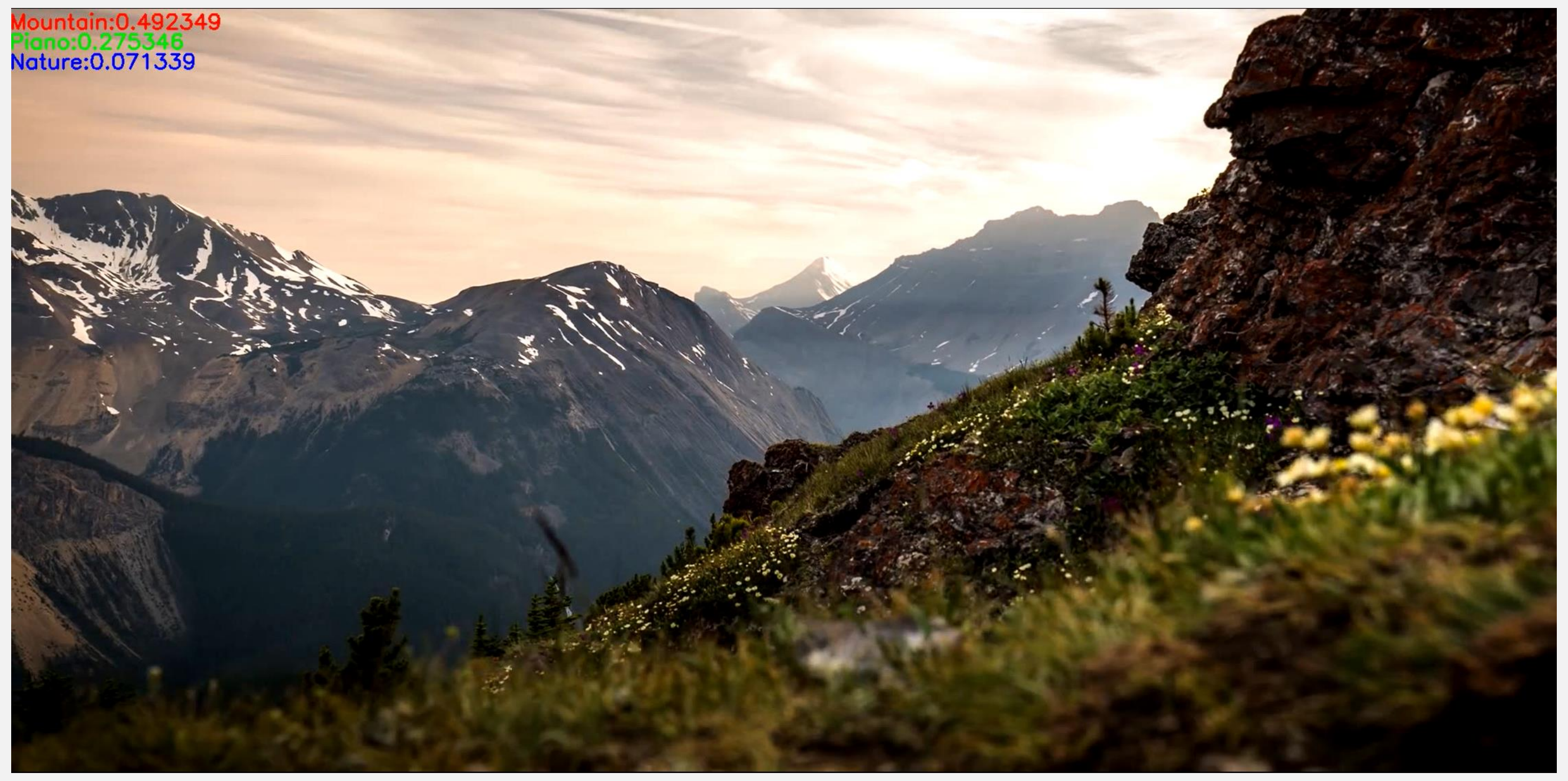
Class	Segments
3	002G:35 002G:40 002G:60
7	002G:35 002G:40 002G:60
8	002G:35 002G:40 002G:60

- Class: Label category of a video.
- 002G: Unique ID of a video.
- 35: Segment Start Time.
- Each Segment is 5 Second's long.

Trained Model

- We trained our model on the YouTube dataset with Deep Bag of frames model.
- We achieved 0.71 accuracy.

Annotated Outputs



Future Plans

- Bookmark Segments of a video.
- Give highlights of a video.