How to turn an hour-long meeting into 5 minutes?

Team: Imane Farhat & Nassim Helou
Faculty advisor: Prof. Colin Fogarty
Company: Ekta Srivastava & Suzana Iacob

KnowNow Platform

KnowNow is McKinsey & Company's internal multimedia platform that hosts and shares knowledge videos between collaborators, including recordings of Zoom meetings

Video content is currently on the rise, and there is a strong demand for short synthesized highlights of knowledge videos on KnowNow

~5400 videos currently on KnowNow
Over **70%** increase in views from 2020 to 2021
Over **67%** of the firm visited the platform in 2020

Problem statement

Collaborators at McKinsey & Company rarely have the time to consume an hour long video to extract the key knowledge

Manual video summaries are created, but these are very time consuming to make and are available for very few videos

Solution: **Create a video summarization tool for KnowNow videos**

Value proposition

~10 mins are spent to summarize key points of a meeting. A consultant holds about 5 meetings in a day.
Saves **~4 hrs/week/consultant**

~1.5 hours to watch and create video synthesis for curators. Approximately 10 meetings/month have curated summaries.
Saves ~15 hrs/month

Project timeline

February - March
Scope definition of the project
Literature research on video summarization techniques

April
Experiment design
Implementation of unsupervised video summarization model on standard datasets from research paper

May
Preprocessing pipeline design including cleaning and punctuation of transcripts
Implementation of transcript summary models using BERT and PageRank

June
Evaluation of summary models on KnowNow transcripts
Segmentation of transcripts using video key scenes and speakers

July - August
Keyframe classification
Integration of all models in final ouput
Evaluation on KnowNow videos

Solution proposition

The objective of this project is to solve this problem by developing capabilities to:
1. **Auto-generate synthesized videos**
2. **Text summaries and key highlights from the video**

Using **state of the art research** to generate static or dynamic video summaries using supervised or unsupervised approaches. The output of these models is a shorter video that contains the key sequences or key frames of the video.

Using **video transcripts and metadata** to capture the key highlights of the video using the following scheme:
- Cleaning and summarizing the transcripts
- Selecting the most relevant frames of the video to accompany the text summary. Since the videos are essentially Zoom recordings, the most relevant frames should be **slides extracted** from the meeting in the video

Datasets

Curated KnowNow videos in mp4 format
The videos used are mostly **long-format Zoom recordings** of meetings (prioritized videos)

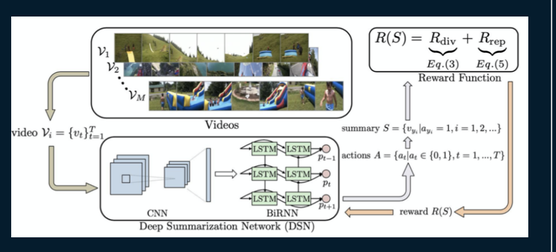
Transcripts of videos generated using existing Speech to Text algorithm in KnowNow

Metadata of videos (speakers in the video)
Selection of most important keyframes of the video generated using a vendor product by Microsoft (Azure Video Indexer)

Experiment 1

After reviewing the literature of this state of the art field, we implemented an approach based on **Reinforcement Learning and CNNs for unsupervised video summarization.**

The Summarization Network selects frames from the original video using a reward function based on **diversity** and **representativeness** of the frames.



The output of this model is a video composed of frames selected from the original video. For evaluation, we implement the model on standard datasets annotated with importance scores of each frame. **The evaluation metric used is the F-1 score** between the frames selected by the model and those selected based on their importance scores.

We achieved 60% F-1 scores on these datasets, which is the target score reached by the authors of the implemented paper.

The model is currently trained on standard datasets and evaluated on KnowNow videos.

Experiment 2

Building a **pipeline to preprocess video transcripts** generated by Speech to Text algorithms.

Cleaning the transcripts → Filtering the transcripts → Fixing punctuation issues

The punctuation algorithm implements a punctuation generating paper using **Bidirectional Recurrent Neural Network** with Attention Mechanisms.

Implementing **extractive** text summaries using **BERT** and **PageRank** which are two Google pre-trained neural network algorithms.

Implementing **abstractive** text summaries using the **BART** model to improve summary quality.

Experiment 3

The video metadata provides insights on:
- The **main speakers** in the video
- The **key scenes** of the video

We use these insights to segment the video and transcript and summarize each segment separately, assuming that **when a speaker or scene changes, the ideas discussed change as well**.

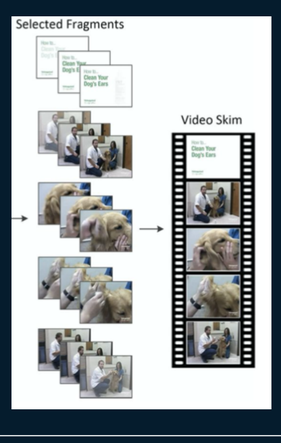
The key frame extraction returns frames of slides and frames of random snapshots of the video. For knowledge Zoom recordings, we assume that the most relevant frames are snapshots of **slides**. Therefore we implement a **classification model to detect whether a frame is a slide or not**.

Feature engineering of the frames to extract **color** indicators and detection of the **presence of text** in the frame. Both are very **strong predictors**.

Implementation of **Random Forest model** for frame classification, resulting in **95% accuracy.**

Results

The output of Experiment 1 is a summarized video: it's a **shorter video containing the main frames of the original video.**
Many people enjoy watching a synthetized video more than reading text, the dynamic element making it easier to remember.



The combination of the **transcript summary from Experiment 2** and the **selected frames corresponding to the most relevant slides** of the video from **Experiment 3** results in creating an **article-like summary** that captures the key highlights of the video.

Impact

17 hours per month saved per consultant

15 hours per month saved per content curator

Well-curated and easily consumable knowledge

Conclusion and Future Directions

Conclusion
- We have laid the foundation for a video summarization model.
- Our transcript preprocessing pipeline can be generalized and used for any KnowNow transcript.
- Our methodologies can be easily modified to incorporate different models and allow for flexibility and experimentation with the output.

Future steps
- Finish building a preprocessing network (GoogleNet or ResNet) to train the model in Experiment 1 on KnowNow videos as well.
- Build infrasctructure needed for data and model hosting in a UI.