Multimodal Abstractive Summarization for Open-Domain Videos

Jindřich Libovický¹, Shruti Palaskar², Spandana Gella³, Florian Metze²

¹Charles University
²Carnegie Mellon University
³University of Edinburgh
Introducing Summarization with How2 Data

- Summarization
  - Present subset of information in a more compact form (maybe across modalities)

- "Description" field
  - 2-3 sentences of meta data: template based, uploader provides
  - "Informative" and abstractive summary of a how-to video
  - Should generate interest of a potential viewer
How2 Dataset

- 2000h of how-to videos
  - 300h for MT
  - 480h for ASR
  - 80K videos
  - Object, Scene and Action features
- Human annotated transcripts
- Meta-data
- Video summaries / descriptions
- Very different topics
  - Cooking, music, sports, etc.

R. Sanabria et al., “How2: A Large-Scale Dataset for Multimodal Language Understanding” ViGIL 2018
Hierarchical Multi-modal Attention

Different Multimodal Models
1. Text-only
2. Video-only
3. Text-and-Video Models
Attention over the Video Features

Talking and preparing the brush

Close-up of brushstrokes **with hand**

Close-up of brushstrokes **no hand**

Black frames at the end
Winning Model in DSTC7 AVSD Track

- Applied this model to the DSTC7 AVSD track
- Goal was to test the effect of pre-training for this task on a large corpus, How2
- Charades dataset: Visual Question Answering type data with audio and summary as extra modalities
- Best performing model in objective and human evaluation
Multimodal Abstractive Summarization for Open-Domain Videos

How2: A Large-Scale Dataset for Multimodal Language Understanding

Dataset: https://github.com/srvk/how2-dataset

Code: https://github.com/lium-lst/nmtpytorch