Multimodal Abstractive Summarization for How2 Videos – Summary (Simple & Clear)

1. What is the Problem?

• The internet has millions of instructional videos (cooking, sports, DIY, etc.).

• Many videos do NOT have useful descriptions or summaries.

• Users find it hard to know what a video contains without watching it fully.

• Goal: Create short, meaningful text summaries for videos using multimodal data (video + audio + transcripts).

2. Why Not Use Normal Text Summarization?

• Traditional summarization is usually applied only to text (like news articles).

• But videos contain important information in BOTH audio and visuals.

• Transcripts alone may miss key details (e.g., visuals like “cutting peppers” or context like “Cuban breakfast”).

3. What is Multimodal Summarization?

• A model learns from:

– Video features (visual understanding)

– Audio transcripts / ASR output (speech)

– Human-written transcripts (when available)

• It combines information from different modalities into ONE final summary.

4. Dataset Used – How2 Dataset

• Large dataset of 2,000+ hours, 70k+ instructional videos.

• Each video has:

– Video content

– Human transcript

– ASR transcript

– 2–3 sentence human-written summary

• Covers many domains: cooking, music, sports, etc.

5. Method Used – Hierarchical Multimodal Attention Model

• Sequence-to-sequence neural model.

• Takes multiple inputs:

– Full transcript (text)

– Action video features (from ResNeXt 3D CNN)

• Uses hierarchical attention:

– Learns what text to focus on

– Learns what video frames are important

• Produces a fluent, readable summary.

6. What Makes This Model Special?

• It fuses both video and text.

• Extracts key actions or context not mentioned in transcripts.

• More complete and accurate summaries than text-only models.

7. New Metric Introduced – Content F1 Score

• ROUGE measures overlap with human summary (word matching).

• But ROUGE does NOT measure whether summary is semantically correct.

• Content F1 measures:

– Semantic similarity

– Whether the model captured the right meaning

• More suited for multimodal summarization tasks.

8. Experiments – Key Findings

• Text-only summaries work well when transcript is long and accurate.

• Video-only models also perform surprisingly well.

• Combined (text + video) model gives the BEST results.

- ASR transcripts reduce accuracy but still give good performance.
- Multimodal model produces the highest ROUGE-L and Content F1.

9. Human Evaluation Results

- Human judges ranked summaries based on:

– Informativeness

– Relevance

– Coherence

– Fluency

- Multimodal model scored highest across almost all categories.

10. Main Contributions of the Paper

- First large-scale study of multimodal abstractive summarization.
- Introduced:

– How2 dataset benchmarking

– Hierarchical multimodal attention model

– New semantic scoring metric (Content F1)

- Proved that combining video + text improves summary quality.

11. Conclusion

- Best results come from using BOTH text (transcripts) and visual features.
- Multimodal summarization helps computers understand videos more like humans.
- Useful for video search, recommendations, and accessibility tools.