Dynamic Video Summarization via Bidirectional RNNs and Transformer Architectures

Course Project: Deep Learning (DSAI 308)

Amr Yasser, Omar Hazem, Ali Ashraf

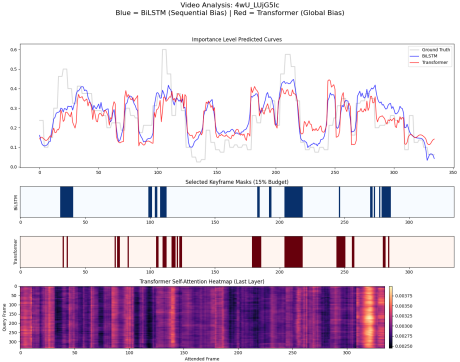
Deep Learning
Supervisor: **Dr. Khaled El-Sayed**

Fall 2025

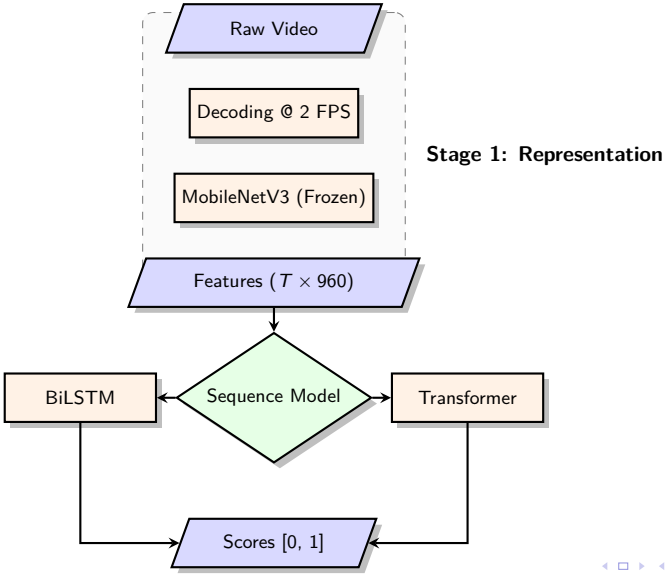
Outline

Motivation & Problem Definition

- **Explosion of Video Content:** Massive growth in platforms requires automated navigation.
- **Video Summarization:** Generating a concise, representative summary (static keyframes).
- **Challenges:**
  - Capturing long-range temporal dependencies.
  - Balancing redundancy vs. information salience.
  - Handling diverse video domains (narratives vs. action).



System Architecture: Two-Stage Pipeline



BiLSTM: Sequential Contextualization

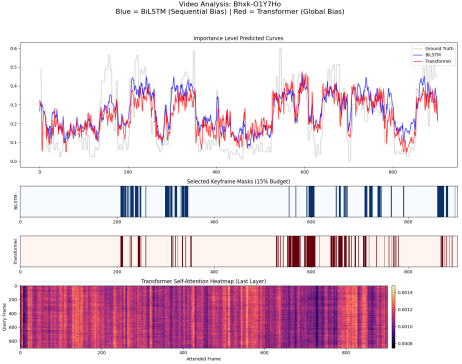
$$h_t = [\overrightarrow{f}(x_t, h_{t-1}); \overleftarrow{b}(x_t, h_{t+1})]$$

Architecture

- 512-dim Bottleneck Projection.
- 2-layer Bidirectional LSTM.
- MLP Regression Head.

Intuition

- Models local flow (past/future dependencies).
- Superior for narrative continuity.



Transformer: Global Dependencies

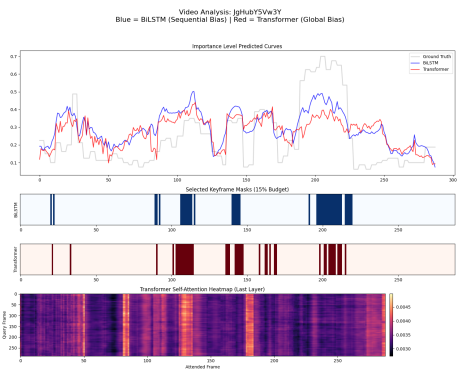
$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Architecture

- Sinusoidal Positional Encoding.
- 4-Head Self-Attention.
- Dense Head for Score Regression.

Intuition

- Non-local relationship modeling.
- Correctly identifies sparse climax events.



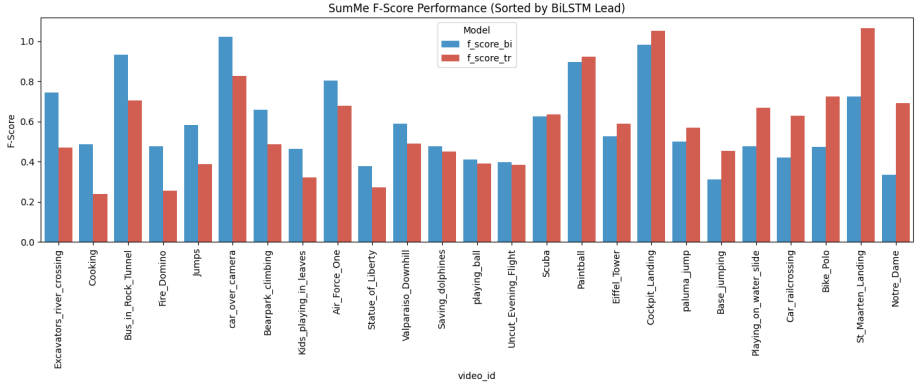
Quantitative Evaluation (TVSum)

Table: Supervised Results on TVSum

Model	Spearman ρ	Kendall τ	MSE	MAE
BiLSTM	**0.531**	**0.362**	0.0142	0.092
Transformer	0.418	0.285	0.0191	0.104

- **Key Finding:** BiLSTM excels at sequential tasks (tutorial, vlogs).
- **Observation:** Transformer shows higher robustness in non-linear action sequences.

Transfer Learning on SumMe Dataset



- Successful zero-shot transfer from TVSum $\rightarrow$ SumMe.
- Models successfully identify climax moments even in unseen domains.

Qualitative Results: Sample Demo Keyframes

Cars Demo Video (T=126)

AI Robot Demo Video (T=33)

Conclusion & Future Work

Summary

- Built an E2E pipeline: decoding, features, modeling, selection.
- BiLSTM is the reliable choice for steady temporal flows.
- Transformers offer better global reasoning for sparse highlights.

Future Directions

- **Multimodal Fusion:** Integrating Audio/Text signals.
- **Unsupervised Learning:** Scaling with SUM-GAN/Contrastive learning.

Questions?

Thank You!