Title: **T**ransformer-based **Q**uery-**A**ware **M**ulti-**V**ideo **S**ummarization

THE UNIVERSITY OF SYDNEY

MOH'D ABU OMAR (520534478), Shiwen Xu (520569045)

COMP5329 Deep Learning

Research Problem

The increasing number of videos uploaded on social media platforms presents a challenge for search engine users who want to find relevant and non-redundant videos in their query results. MVS (Multi Video summarization) aims at getting informative summary frames from a list of videos frames. Deep Query-Aware MVS can be considered a state-of-the-art-bench-mark reinforcement learning model (As shown in *Figure 1*), with LSTM as its backbone component. [1]

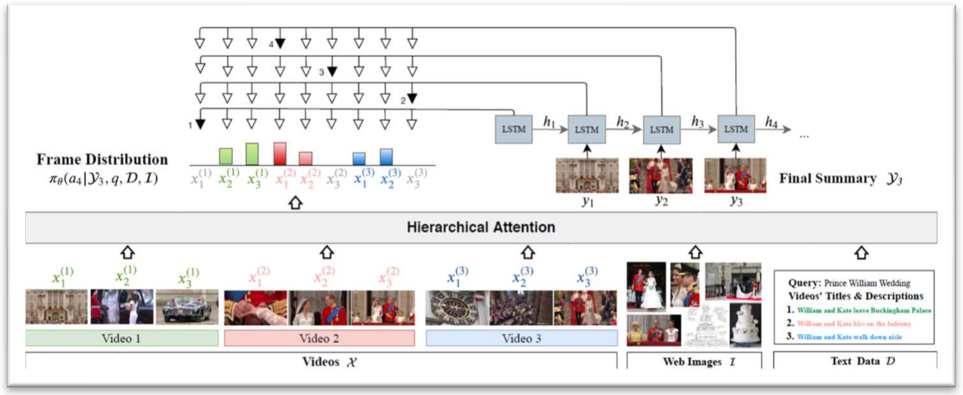
Frame Distribution
$\pi_\theta(a_4|\mathcal{Y}_3, q, D, \mathcal{I})$
LSTM
h_1
h_2
h_3
h_4
Final Summary $\mathcal{Y}_3$
y_1
y_2
y_3
Hierarchical Attention
Video 1
Video 2
Video 3
Videos $\mathcal{X}$
Web Images $\mathcal{I}$
Query: Prince William Wedding
Videos' Titles & Descriptions
1. William and Kate leave Buckingham Palace
2. William and Kate kiss on the balcony
3. William and Kate walk down aisle
Text Data D

Figure 1: DeepQAMVS

Methodology

Due to their parallelization and attention mechanism, Transformers are more efficient in getting long-range relations between the input sequences, with less training time required compared to LSTM. TQAMVS (Transformer-based QAMVS) is suggesting a new method based on DeepQAMVS, with Transformers [5] as a backbone component. (As shown in *Figure 3*)

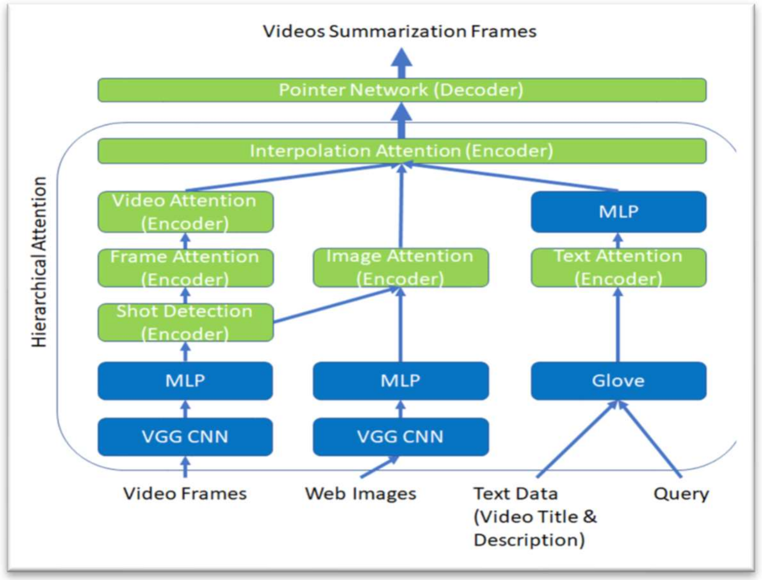
Videos Summarization Frames
Pointer Network (Decoder)
Interpolation Attention (Encoder)
Hierarchical Attention
Video Attention (Encoder)
Frame Attention (Encoder)
Shot Detection (Encoder)
Image Attention (Encoder)
MLP
Text Attention (Encoder)
MLP
MLP
Glove
VGG CNN
VGG CNN
Video Frames
Web Images
Text Data (Video Title & Description)
Query

Figure 2: TQAMVS Components

- Glove is used to extract text embeddings. [4]
- VGG CNN is used to extract visual embeddings. [3]
- MLP is used to scale up/down the dimensions. [6[
- Transformer Encoders are used for the attentions.
- Pointer network is based on Transformer Decoder to get the summary.

Plan

The plan has been split into different tasks. (As shown in *Figure 3,* completed tasks marked in green).

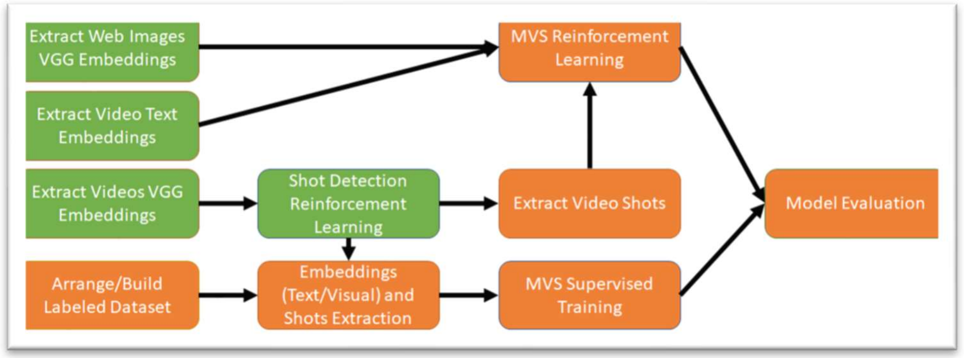
Extract Web Images VGG Embeddings
Extract Video Text Embeddings
Extract Videos VGG Embeddings
Arrange/Build Labeled Dataset
Shot Detection Reinforcement Learning
Embeddings (Text/Visual) and Shots Extraction
MVS Reinforcement Learning
Extract Video Shots
MVS Supervised Training
Model Evaluation

Figure 3: Plan Tasks

- Embeddings Extraction to be done in advance to save the training time.
- Shot detection reinforcement learning uses representativeness reward with TVsum [2] dataset. (Sample frames are shown in *Figure 4*)

Figure 4: Sample TVsum video frames for changing vehicle tire

- MVS reinforcement learning uses TVsum dataset with manually downloaded web images (samples in *Figure 5*), and has multiple rewards (diversity, representativeness, query adaptability and temporal coherence).

Figure 5: Sample manually downloaded web images for changing vehicle tire

- MVS supervised training dataset was unavailable, and may need to be created manually to proceed with the related tasks.

Discussion

- Research scope needs to align with the related constraints (e.g., time), as some tasks were incomplete.
- Feature extraction did not consider MLP training to learn the different images/frames, so that it is expected to lose some distinctive features while scaling down the visual features dimension.
- Interpolation Attention combines multiple attentions using learned weights (i.e., attention of the attentions).
- Combining Reinforcement Learning with Supervised Training is expected to help further optimize the model.

Results

Shot detection results were of little diversity or representativeness. (As shown in *Figure 6*) This is aligned with the expected lost distinctive features for scaling down.

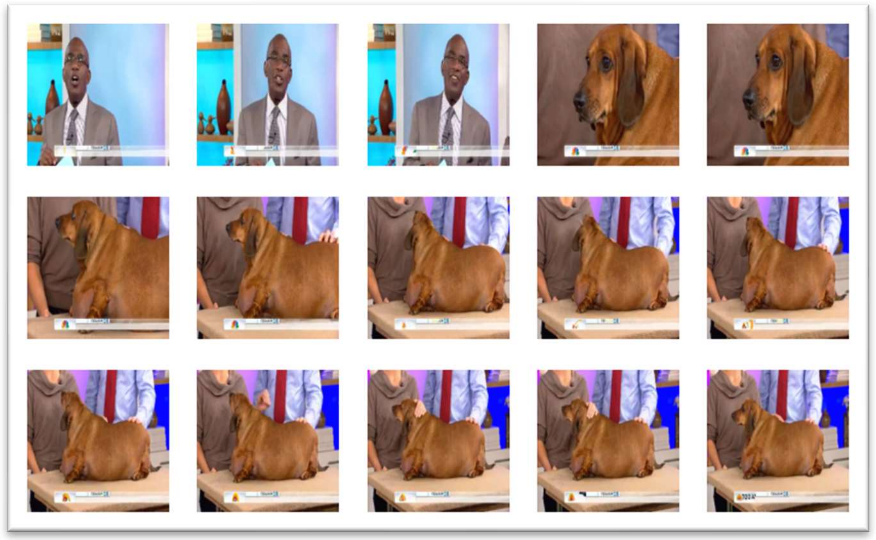

Figure 6: Sample shot detection.

Even though the research is incomplete, the new method is expected to outperform DeepQAMVS [1], with higher F1-score/accuracy, and less training time.

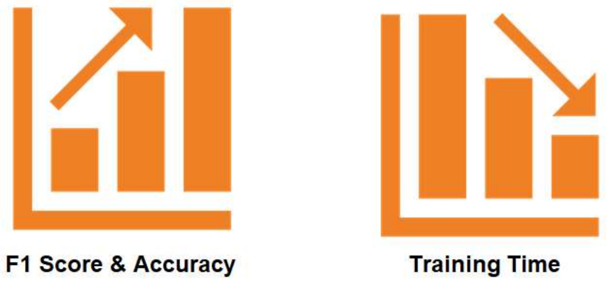

Figure 7: Expected TQAMVS vs DeepQAMVS Result

References

[1] Messaoud, S., Lourentzou, I., Boughoula, A., Zehni, M., Zhao, Z., Zhai, C., & Schwing, A. G. (2021). DeepQAMVS: Query-Aware Hierarchical Pointer Networks for Multi-Video Summarization. ArXiv:2105.06441 [Cs]. https://arxiv.org/abs/2105.06441

[2] Song, Y., Vallmitjana, J., Stent, A., & Jaimes, A. (2015). TVSum: Summarizing Web Videos Using Titles. Openaccess.thecvf.com. https://openaccess.thecvf.com/content_cvpr_2015/html/Song_TVSum_Summarizing_Web_2015_CVPR_paper.html

[3] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," undefined, 2014. https://www.semanticscholar.org/paper/Very-Deep-Convolutional-Networks-for-Large-Scale-Simonyan-Zisserman/eb42cf88027de515750f230b23b1a057dc782108 (accessed Aug. 09, 2020).

[4] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, doi: https://doi.org/10.3115/v1/d14-1162.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, Aidan N, Kaiser, L., & Polosukhin, I. (2017b). Attention Is All You Need. ArXiv.org. https://arxiv.org/abs/1706.03762

[6] Xu, C. (S1, 2023b). *Deep Learning COMP5329, Lecture-2, W2-Multilayer Neural Network* [PDF] Retrieved from https://canvas.sydney.edu.au/courses/48406/files/29762528?wrap=1