Applications of Convolutional Neural Networks on Multimodal Video Summarization

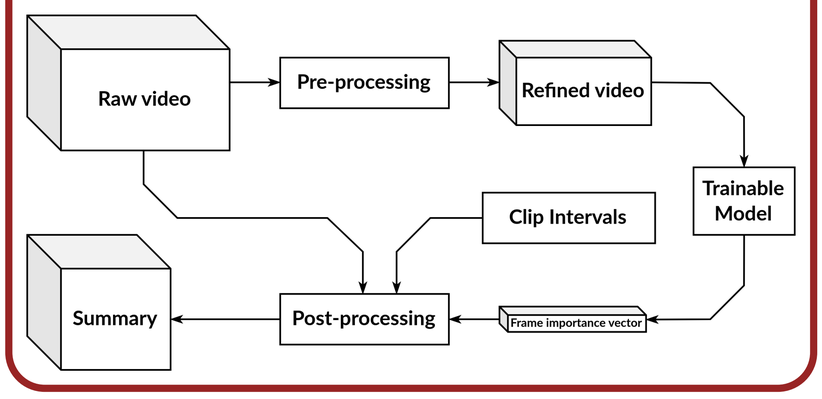
Vasileios Savvas Papagrigoriou, Apostolos Dimoulakis

vpapagr@csd.auth.gr, adimoulak@csd.auth.gr Department of Informatics, Aristotle University of Thessaloniki

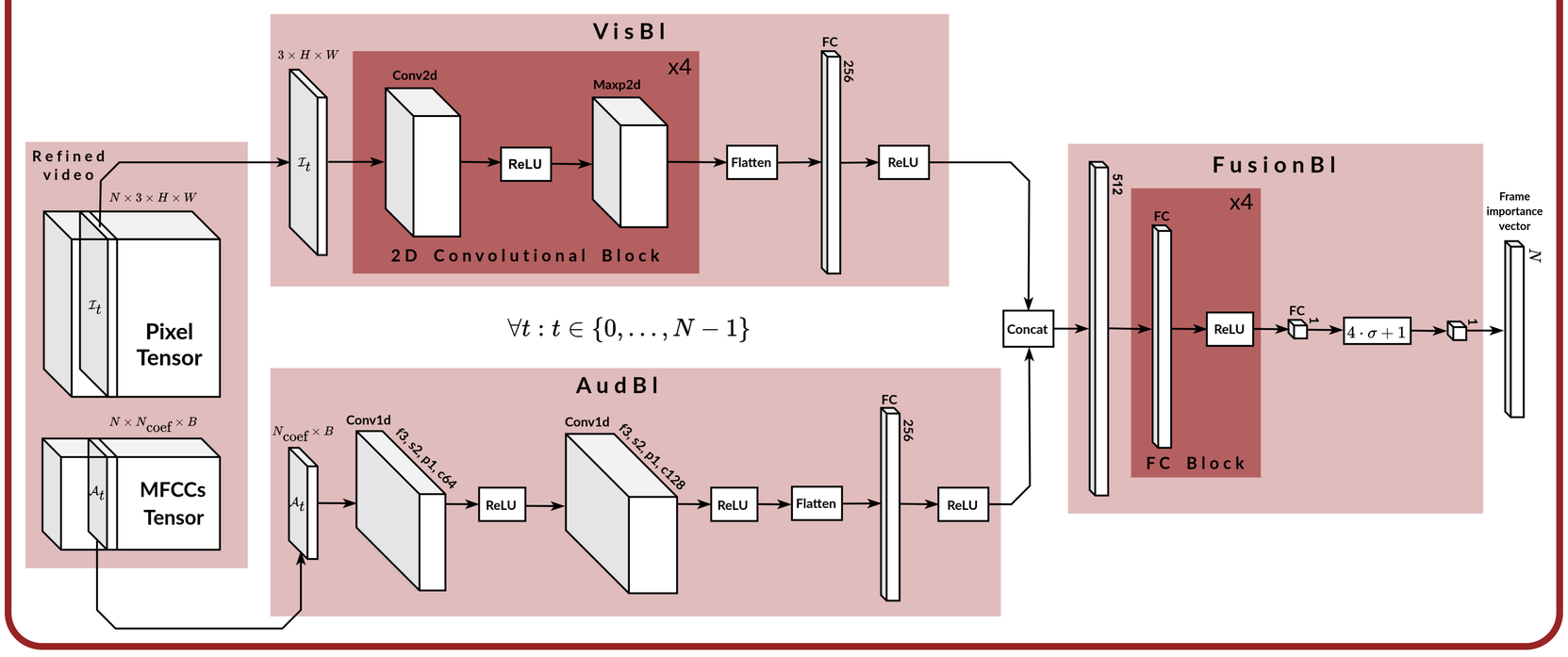
1. Overview

High length videos can sometimes contain segments that lack contextual relevance. Video summarization is the task of reducing a video to its most important segments. In our work, we have performed an ablation study, by investigating and comparing 4 lightweight video summarization models. Our analysis focuses on simple convolutional neural networks capable of modeling the importance of individual video frames. To perform our study we have utilized supervised learning that was relied on a benchmark dataset. Our current work is a reliable starting point for the task of video summarization. Our project is available on **GitHub**:

2. Pipeline



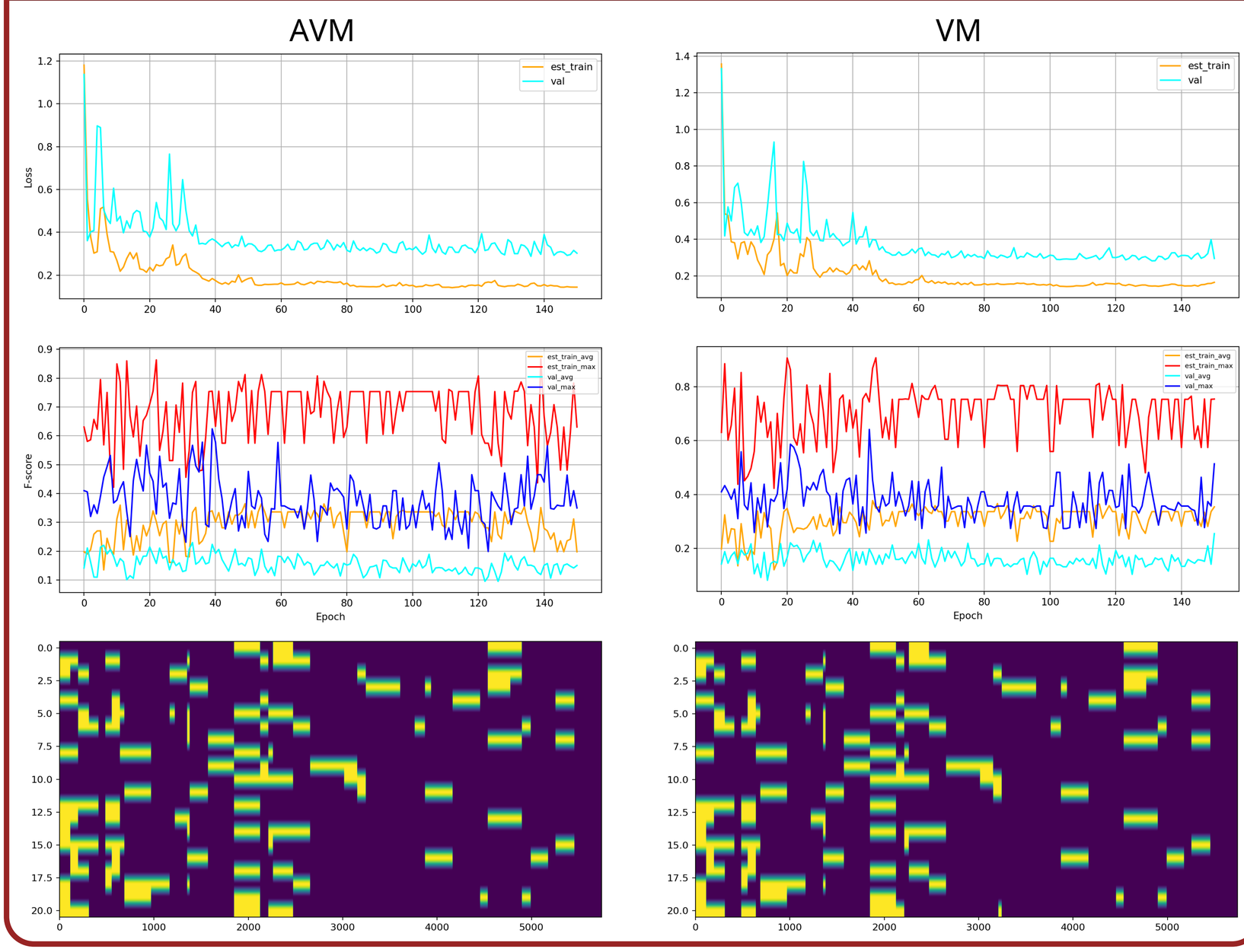
3. Architecture of AVM



5. Ablation Study Table

Model Name	Output (#)	Loss Function	Audio	Train F-score Avg	Train F-score Max	Test F-score Avg	Test F-score Max
AVM	$4 \cdot \sigma + 1$ (1)	MSE	True	0.365	0.812	0.149	0.281
VM	$4 \cdot \sigma + 1$ (1)	MSE	False	0.377	0.869	0.172	0.454
CAVM	Softmax (5)	Cat. Cross Entropy	True	0.263	0.708	0.142	0.410
CVM	Softmax (5)	Cat. Cross Entropy	False	0.198	0.630	0.142	0.410

6. Ablation Study Figures



4. Post-processing

This approach was inherited by [1]. The task is to find which clips will be selected to be kept and which will be dropped, targeting a summary with $r := 15\%$ of the original video. The underlying process takes into account the clip intervals, their lengths (w_j) and their inferred importances (c_j). Specifically the task of video summarization is now reduced to this Knapsack 0-1 problem:

> Find a value $x_j \in \{0, 1\}$ for every interval with index j, that maximizes
>
> $$\sum_j (x_j \cdot c_j) \text{ satisfying } \sum_j (x_j \cdot w_j) \leq \lfloor r \cdot N_{\text{raw}} \rfloor$$

8. Conclusions

1. Classifier CNNs underperform
2. Regressor CNNs perform better
3. Gap between train and test losses; observed data shift and overfitting
4. Audio did not contribute to performance

7. Discussion & Comparison

From the above table it can be observed that the classifier models CAVM and CVM have obviously failed to improve. In contrast, the regression models AVM and VM have been improved during the training session. Considering the loss diagrams of the above figure, we can see that the latter have gradually improved throughout the first 40 epochs. The discrepancy between the train and validation/test losses is an obvious sign of data shift because they have almost identical trends. Since the train metrics are all consistently worse on the test set, we can accept that the model overfits. To compare AVM with VM, audio did not seem to contribute to the performance of our two regressors. Our models are outperformed by the state of the art models such as [2].

10. References

[1] Evlampios Apostolidis. Summarizing videos using concentrated attention. ICMR '22, page 407–415, New York, NY, USA, 2022. Association for Computing Machinery.

[2] Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. Video summarization with attention-based encoder-decoder, 2018.