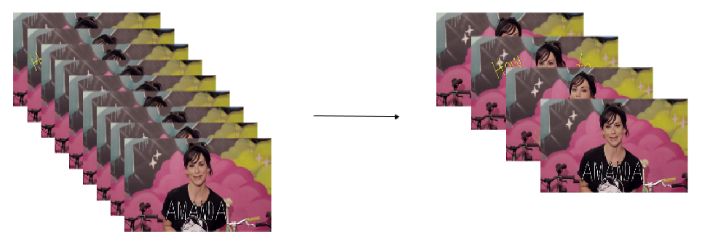
Video Summarization

Manu Agarwal, Kunal Lad
University of Texas at Austin

Problem Statement and Motivation

Generate a summary of the most vital parts of the video. It is motivated by the success of Sequential Determinantal Point Process (DPP) for supervised video summarization. Also inspired by sequence to sequence video to text (s2vt).

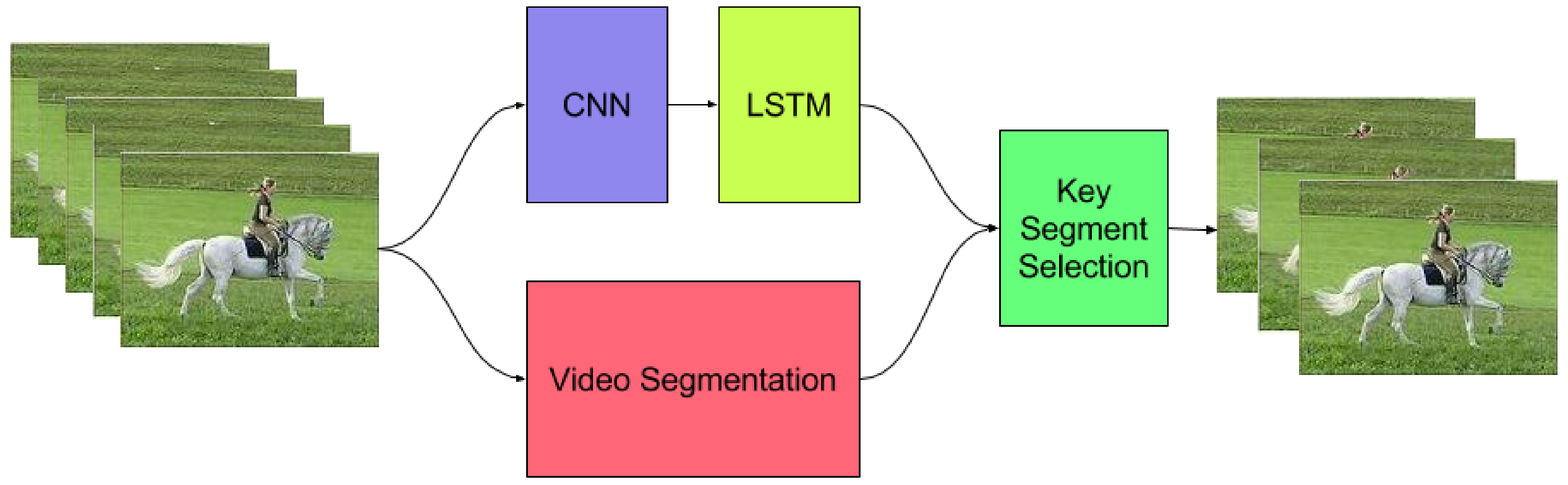


Related Work

Most of the work has been focused on generating summaries using key frame detection based on some selection criteria. Some works use key segments rather than key frames. Only recently have people started modeling temporal information of videos for this task.

Our Approach

We use a combination of Convolutional Neural Network(CNN) and Long Short Term Memory (LSTM) modules. LSTM exploits the temporal information contained in the video.



Dataset

TVSum50 dataset: 50 videos of different genres collected from Youtube. Contains 20 annotations of shot-level importance scores per video

Importance Score: 3

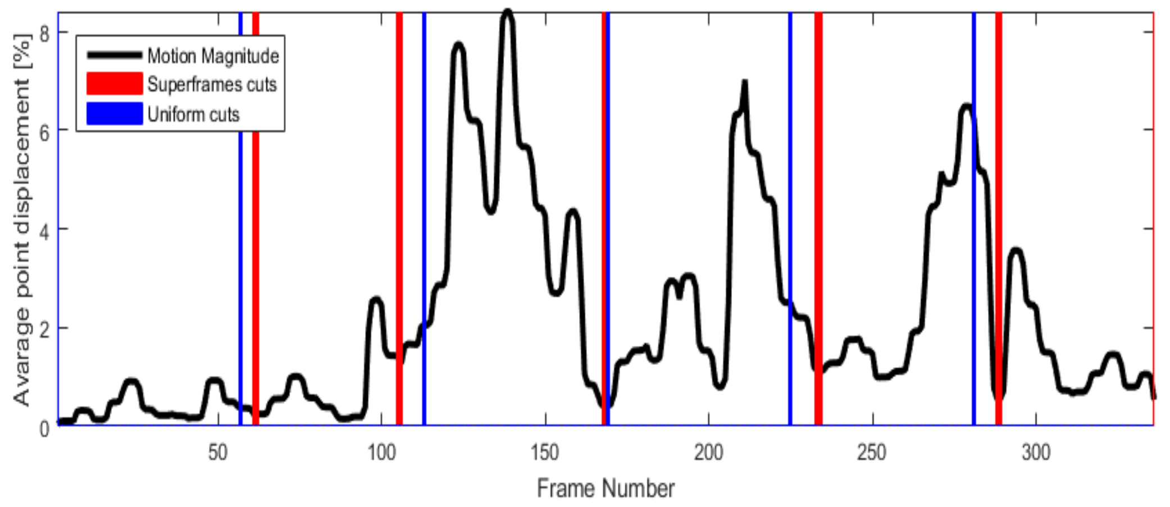
Importance Score: 4

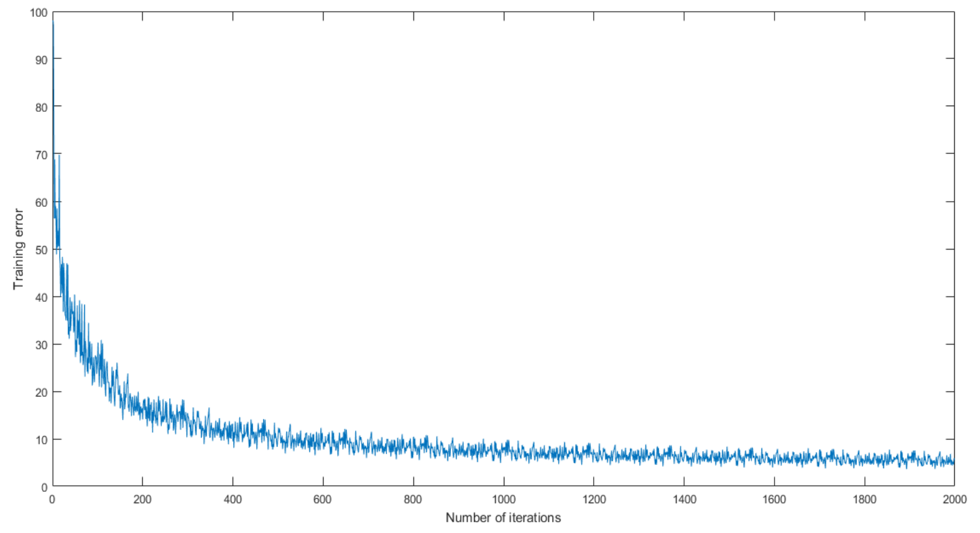
Importance Score: 1

Methodology

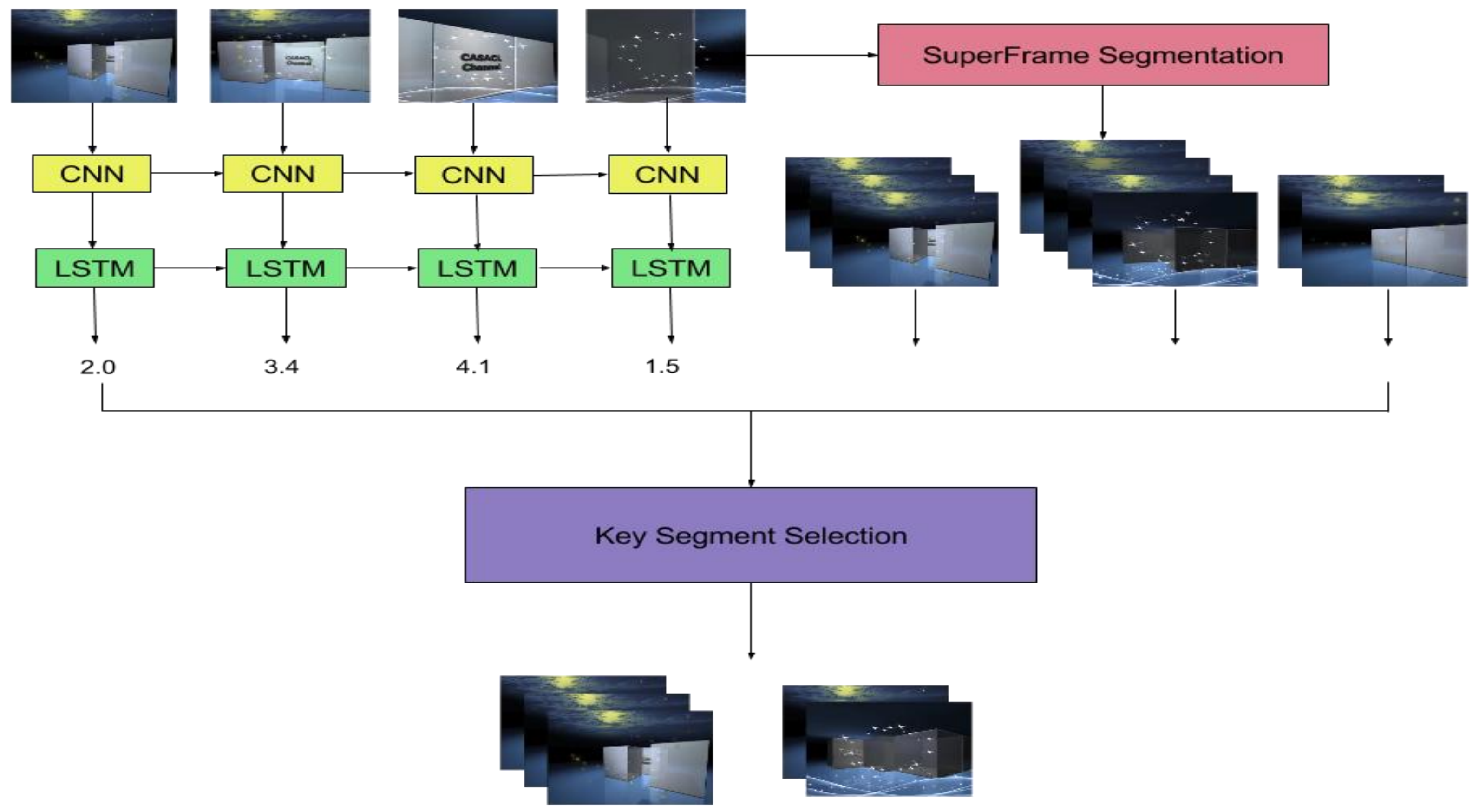
Superframe Segmentation

We will be using Superframe Segmentation framework of Gygli et al. for segmenting the videos. It uses motion cues and principles from video editing theory to compute segment boundaries which are visually aesthetic. The main advantage of this approach is that it works well even for raw videos.



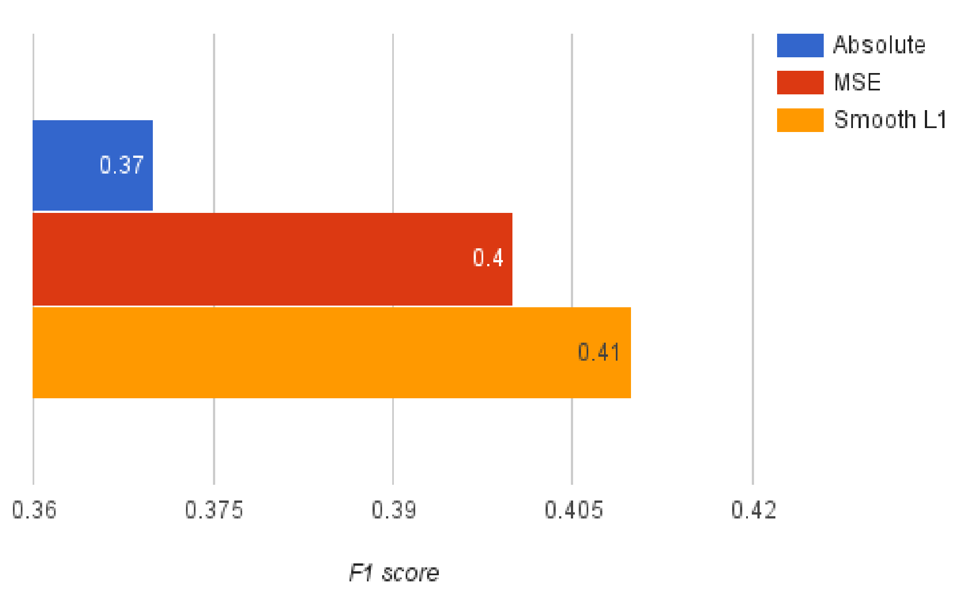


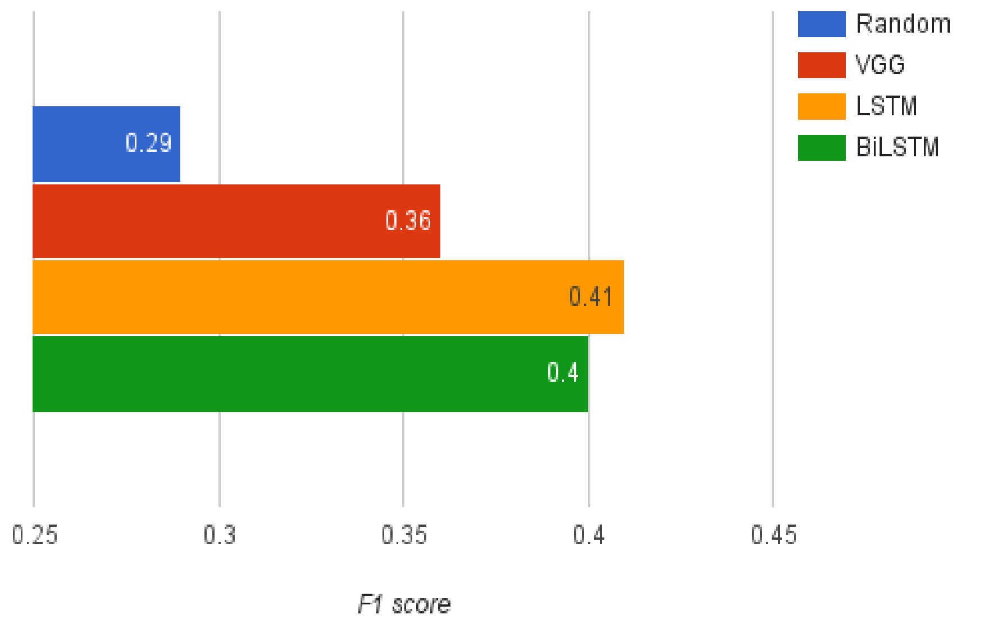
Architecture



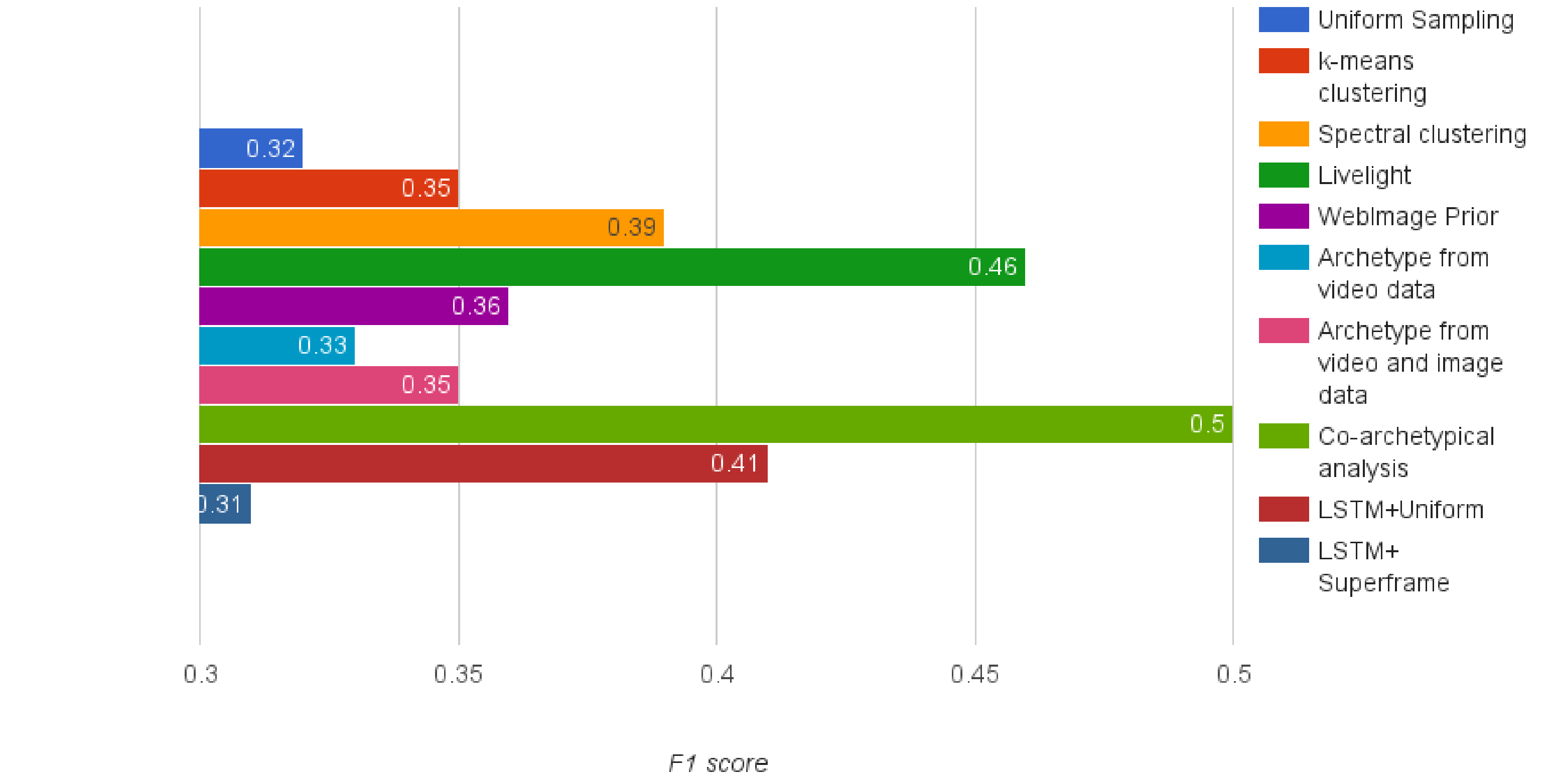
- ▪Video frames are propagated through a pretrained CNN (VGG) network and features from last fully connected layers are extracted.
- ▪CNN features are used to train an LSTM network which predicts frame importance scores.
- ▪Visually aesthetic video segments are obtained by using Superframe segmentation framework [M. Gygli et. al]
- ▪Finally Knapsack is used to generate a summary of key superframe segments on the basis frame scores computed by LSTM.

Results





We use F1 score as the metric to compare the quality of the generated video summary against ground truth annotations. Results show that Smooth L1 criterion outperforms Absolute and MSE loss criterions.



All our models (VGG, LSTM, BLSTM) significantly outperform the baseline (random segmentation with random frame scoring) and beat existing techniques other than Livelight and Co-Archetypical models.

Conclusion

This approach demonstrates the effectiveness of LSTMs for Video Summarization. Availability of larger dataset of annotated videos along with better segmentation techniques will lead to better summary generation.

References

[1] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In ECCV, 2014
[2] A. S. A. J. Yale Song, Jordi Vallmitjana. Tvsum: Summarizing web videos using titles. 2015. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)