Evaluation of Video Summarization Using FCSN on Action Localization Datasets

Paul Frölke p.r.frolke@student.tudelft.com

Supervised by:
Ombretta Strafforello
dr. Seyran Khademi

TU Delft

Video Summarization

Summarizing an input video by selecting the most informative parts.

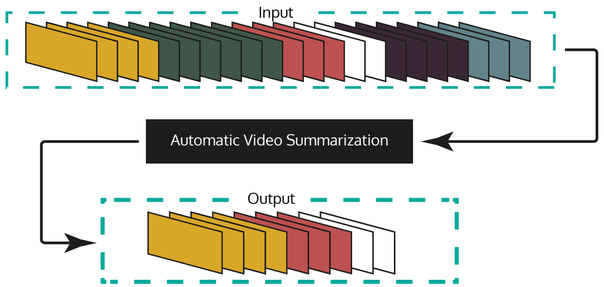

Fig 1: Illustration of video summarization

Problem: subjectivity in training data. What is a good summary?

Action localization

Fig 2: Example video with labels from the Breakfast Actions [1] dataset.

Hypothesis: using the action localization segements when hand labeling summaries reduces subjectivity.

FCSN

Fully convolutional sequence network

SUM-FCN [2] supervised learning

SUM-FCN$_{unsup}$[2] unsupervised learning

Output: key-frame summary

Fig 3: Example key-frame summary, some isolated frames are selected

Evaluation metrics

F1-score: agreement between generated and reference summaries

Rank-order correlation: comparison of ranking frames by their importance scores

Key-frame to imp. score

Problem: predicted prob. of key-frame is does not correspond to importance scores

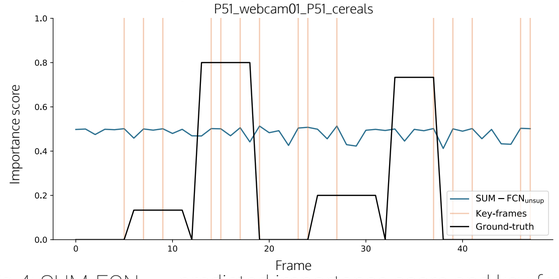

Fig 4: SUM-FCN$_{unsup}$ predicted importance score and key-frames

Result Comparison

Dataset	Method	F1-score	Kendall's τ	Spearman's ρ
TVSum	SUM-FCN	56.5	0.006	0.009
	SUM-FCN$_{unsup}$	52.9	0.009	0.013
	Random	56.4	0.000	0.000
	Human	53.8	0.177	0.204
SumMe	SUM-FCN	30.9	-0.003	0.004
	SUM-FCN$_{unsup}$	28.3	0.000	-0.011
	Random	18.7	0.000	0.000
	Human	31.1	0.202	0.213
Breakfast	SUM-FCN	31.4	0.024	0.032
	SUM-FCN$_{unsup}$	20.1	-0.020	-0.021
	Random	21.4	0.000	0.000
	Human	43.2	-	-

Table 1: Scores achieved on brenchmark datasets

Type	Model	F1-score	Kendall's τ	Spearman's ρ
Supervised	VASNet	67.3	0.037	0.045
	DSNet (Anchor-based)	64.4	0.090	0.106
	DSNet (Anchor-free)	60.0	0.056	0.078
	SUM-FCN	31.4	0.024	0.032
Unsupervised	SUM-GAN-AAE	51.4	-0.030	-0.030
	SUM-FCN$_{unsup}$	20.1	-0.020	-0.021

Table 2: Comparison of other methods on the Breakfast Actions dataset [1].

Main Conclusions

- SUM-FCN performs better on Breakfast Actions compared to other benchmarks.
- SUM-FCN$_{unsup}$ performs worse.
- Compared to other models, SUM-FCN performs slighly worse and SUM-FCN$_{unsup}$ slightly better.
- However, rank-correlation evaluation is not suited for the FCSN models.

[1] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, p 780–787, 2014.

[2] Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, p 347–363, 2018.