Evaluation of Video Summarization using DSNet and Action Localization Datasets

Groenewegen, D. (d.h.e.groenewegen@student.tudelft.nl)

Background

Subjectivity of human annotations can cause discriminative labels when used for video summarization

A proposed solution is using action localization datasets

A supervised method for video summarization is DSNet framework.

Framework has two approaches, anchor-based and anchor-free

Sample of Breakfast Actions dataset [1]

Methodology

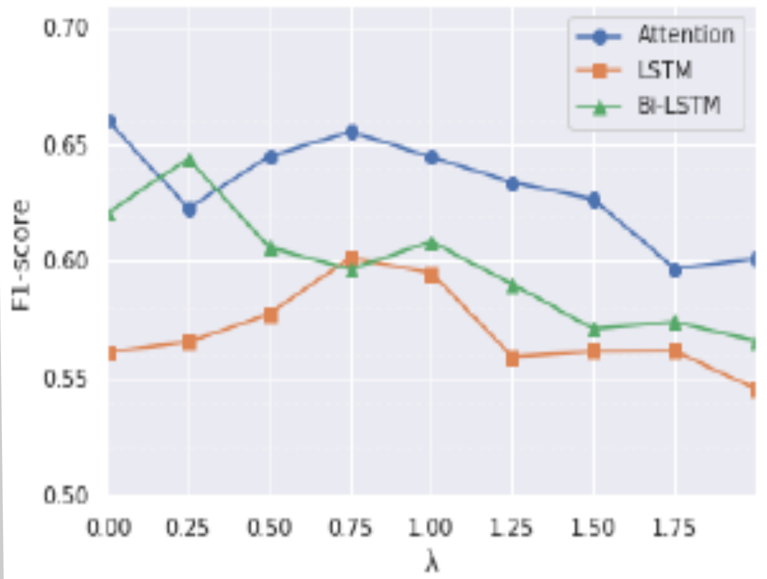
Verify the previous research with TVSum and SumMe datasets

Evaluate approaches with Breakfast Actions dataset using F1-score

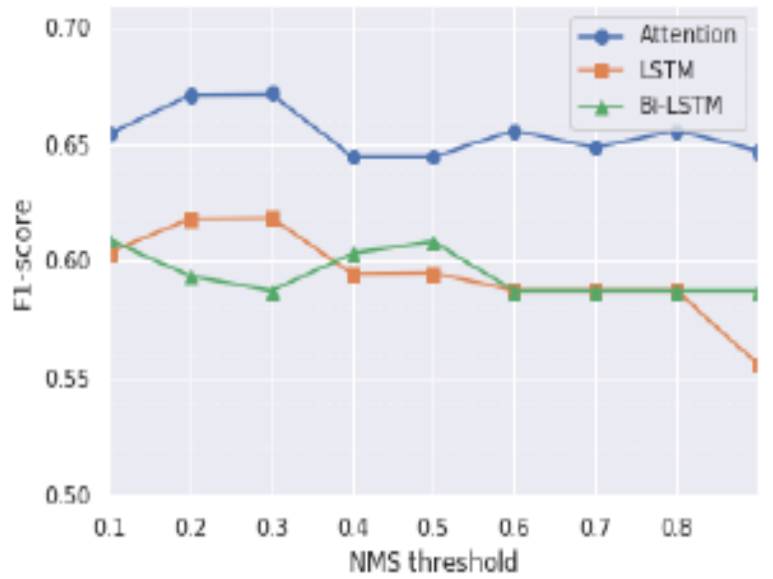
Evaluate the framework with correlation coefficients

Results

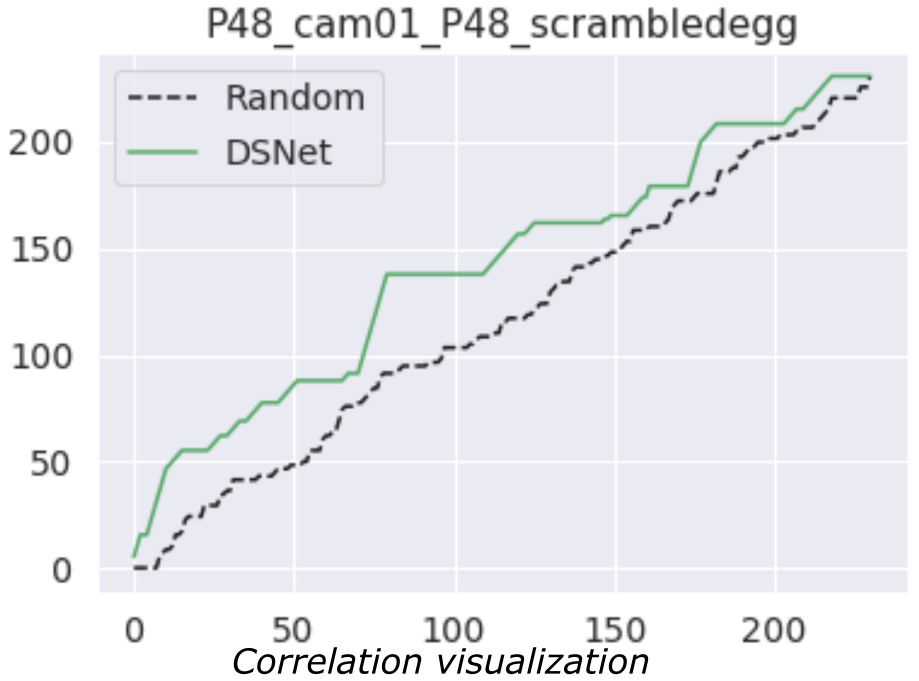
Anchor-Based



Parameter analysis of loss function

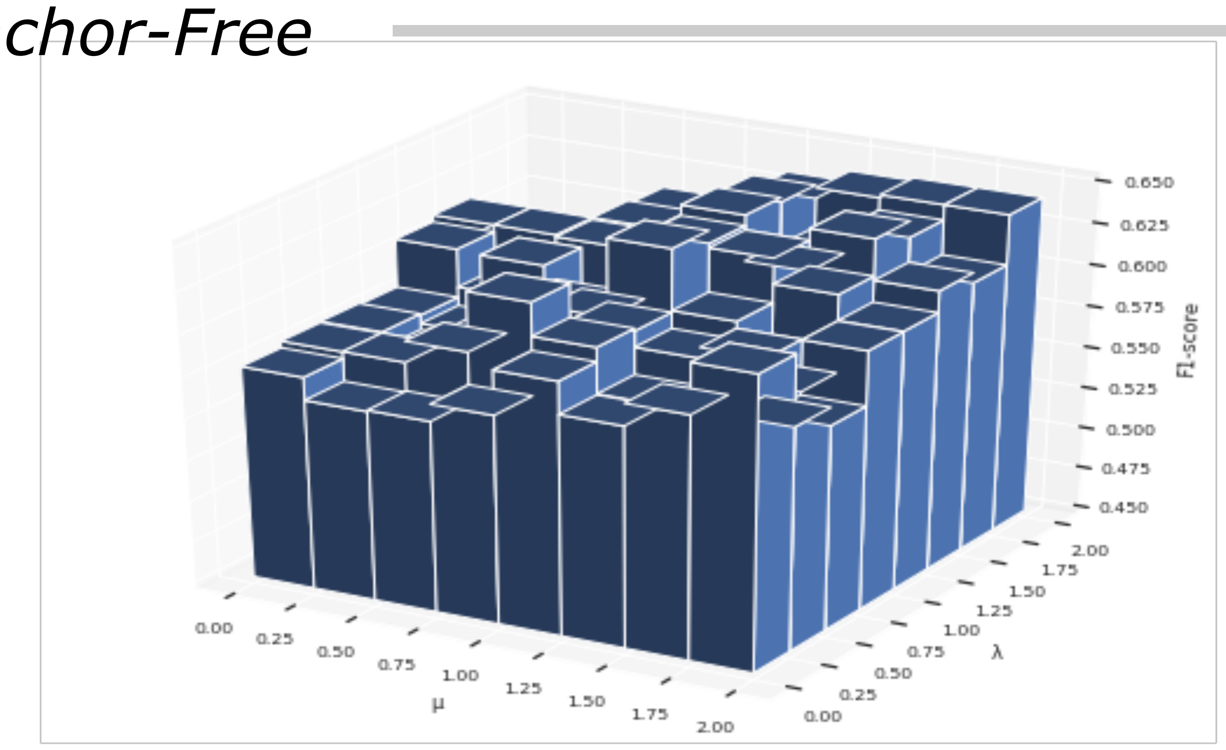


Parameter analysis of NMS threshold

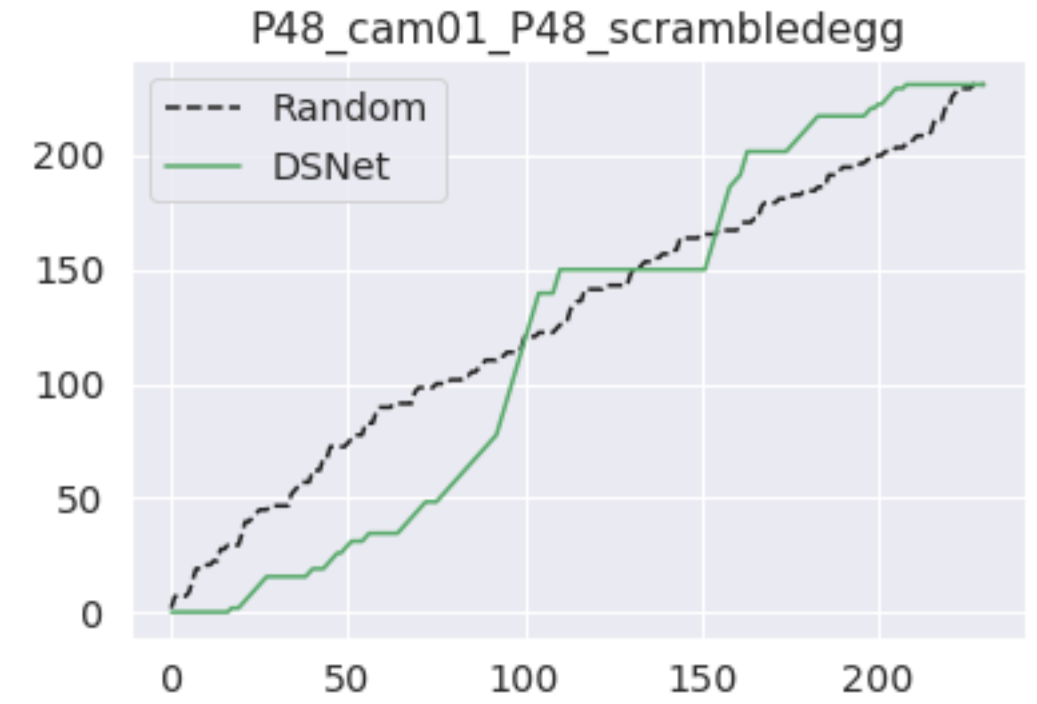


Correlation visualization

Anchor-Free



Parameter analysis of loss functions



Correlation visualization

Correlation coefficients

Dataset	F1 score	Spearman's ρ	Kendall's τ
TVSum (AB)	0.622	0.285	0.198
TVSum (AF)	0.596	0.197	0.276
SumMe (AB)	0.503	0.035	0.041
SumMe (AF)	0.508	0.048	0.062
Breakfast (AB)	0.6446	0.106	0.090
Breakfast (AF)	0.6003	0.078	0.056

Correlation coefficients comparison between different datasets

Conclusion

Breakfast Actions dataset barely improves the accuracy of video summarization

Correlation coefficient lower compared to TVSum

Similar results among the two approaches

[1] H. Kuehne, A. B. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In Proceedings of Computer Visionand Pattern Recognition Conference (CVPR), 2014.