Video Summarization with a Dual Attention Capsule Network

Hao Fu[1], Hongxing Wang[1], Jianyu Yang[2]

[1]Chongqing University, China
[2]Soochow University, China

Problem

Generating a compact and non-redundant summary for a given video without missing significant information.

Contributions at a Glance

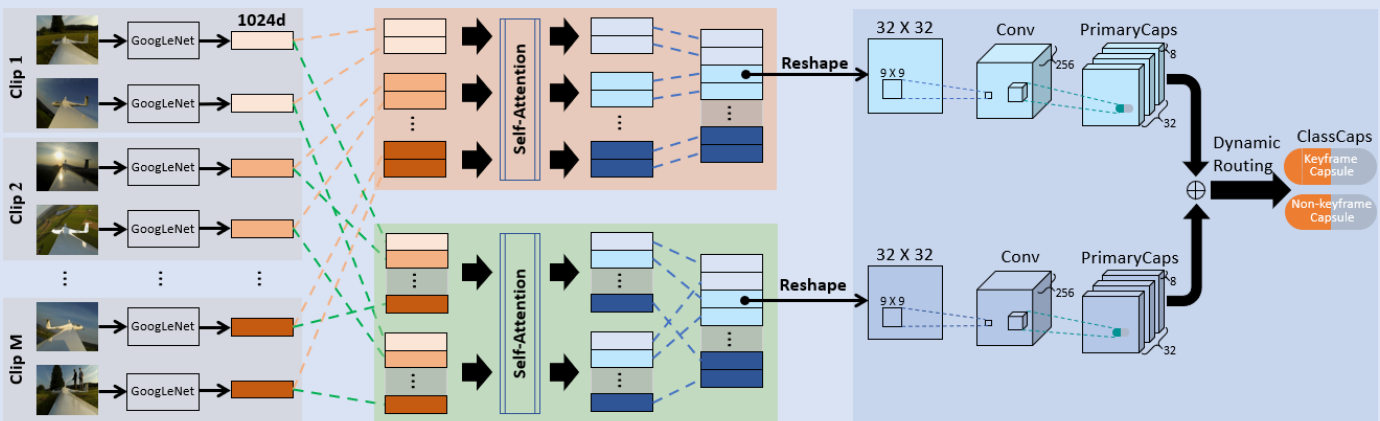
- We propose a novel dual attention capsule network model, which can effectively incorporate the short- and long-term temporal dependencies among video frames for summarization.
- Our proposed video summarization is parallelizable, which can easily handle longer-term dependencies among video frames than the RNN/LSTM-based approaches.
- Experimental results show that our proposed method owns stronger learning ability, and is competitive with existing state-of-the-art methods.

Method

Feature Extraction
- Extractor: GoogLeNet
- Divided frames into M clips
- Each clip is viewed as a local block

Dual Attention Feature Refinement
- Self-attention mechanism
- Learn the short- and long-term dependencies within and between clips

Two-Stream Capsule Network Learning
- Later fusion mechanism
- Scalar → Vector
- Margin loss



Feature Extraction of Video Frames | Dual Attention Feature Refinement Within and Between Clips | Two-Stream Capsule Network for Feature Fusion and Frame Labeling

Experimental Results

Ablation study of our method

Method	SumMe	TVSum
Ours-local	45.2	58.3
Ours-global	45.4	58.5
Ours-fc	46.6	58.7
Ours	**47.5**	**59.4**

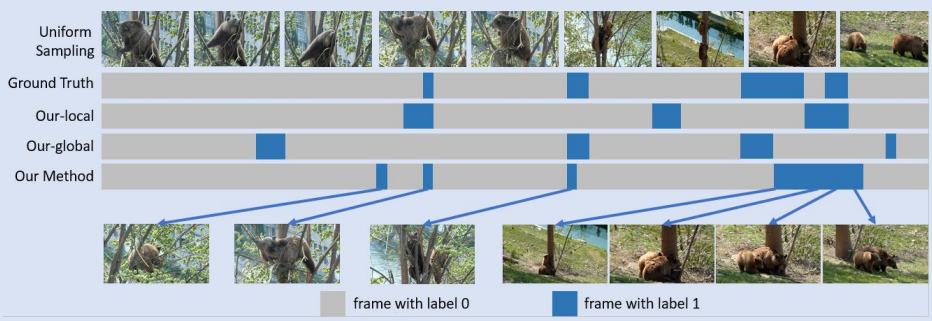
τ and ρ results compared with SOTA

Dataset	SumMe		TVSum	
Metric	Kendall's τ	Spearman's ρ	Kendall's τ	Spearman's ρ
Random	0.000	0.000	0.000	0.000
DR-DSN$_{sup}$ [7]	0.034	0.041	0.025	0.039
dppLSTM [5]	0.040	0.049	0.042	0.055
Ours	**0.063**	**0.059**	**0.058**	**0.065**

F-score(%) results compared with SOTA

Method	SumMe			TVSum		
-	C	A	T	C	A	T
dppLSTM [5]	38.6	41.6	40.7	54.2	57.9	56.9
SUM-GAN$_{sup}$ [6]	41.7	43.6	-	56.3	**61.2**	-
DR-DSN$_{sup}$ [7]	42.1	43.9	42.6	58.1	59.8	58.9
SASUP [35]	45.3	-	-	58.2	-	-
CSNet$_{sup}$ [21]	**48.6**	48.7	44.1	58.5	57.1	57.4
Ours	47.5	**49.3**	**45.2**	**59.4**	59.8	**59.2**

An example summary generated by our approach on the "Bearpark climbing" video



References

- K. Zhang, W. Chao, F. Sha, and K. Grauman, Video summarization with long short-term memory, ECCV'16.
- B. Mahasseni, M. Lam, and S. Todorovic, Unsupervised video summarization with adversarial LSTM networks, CVPR'17.
- K. Zhou, Y. Qiao, and T. Xiang, Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward, AAAI'18.
- C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, Going deeper with convolutions, CVPR'15
- Y. Jung, D. Cho, D. Kim, S. Woo, and I. Kweon, Discriminative feature learning for unsupervised video summarization, AAAI'19.
- H. Wei, B. Ni, Y. Yan, H. Yu, X. Yang, and C. Yao, Video summarization via semantic attended networks, AAAI'18.