Ping LI, Chao TANG, Xianghua XU, 2021. Video summarization with a graph convolutional attention network. *Frontiers of Information Technology & Electronic Engineering*, 22(6):902-913. https://doi.org/10.1631/FITEE.2000429

# Video summarization with a graph convolutional attention network

**Key words:** Temporal learning; Self-attention mechanism; Graph convolutional network; Context fusion; Video summarization

Corresponding author: Ping LI
E-mail: patriclouis.lee@gmail.com
ORCID: https://orcid.org/0000-0002-8515-7773

# Motivation

1. Previous methods do not fully consider the local and global temporal relations among video frames, which play important roles in identifying representative frames.

2. The intrinsic structure of frame samples is not well respected, failing to accurately encode the semantic relations within key frames.

3. Graph convolution networks have shown impressive performance enhancement in vision applications.

# Main idea

1. A graph convolutional attention network is designed to capture the dynamic nature of video with more discriminative and robust video summarization ability.

2. An embedding learning part is implemented by going through two parallel branches, i.e., a temporal branch and a graph branch.

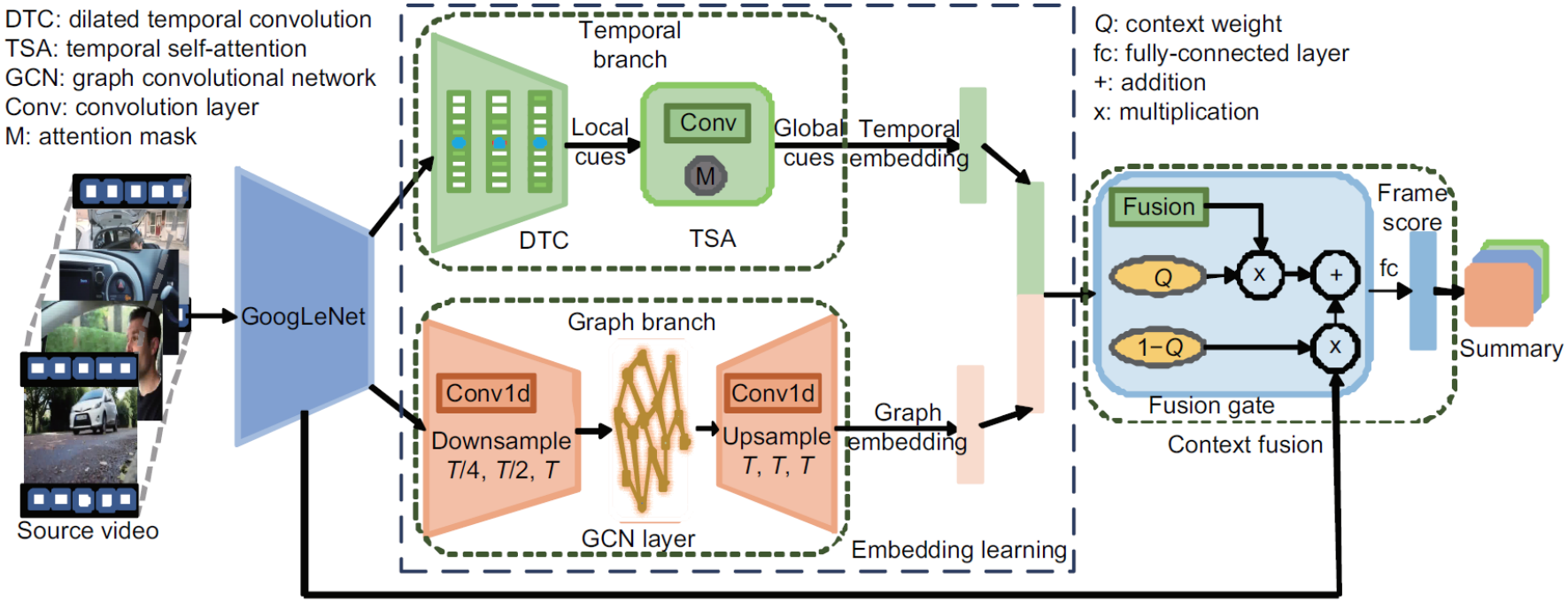
3. A fusion gate is devised to simultaneously consider the temporal context among adjacent frames and the graph context from the entire video.

# Method

1. A novel video summarization approach based on embedding learning and context fusion, named the graph convolutional attention network (GCAN), is proposed.

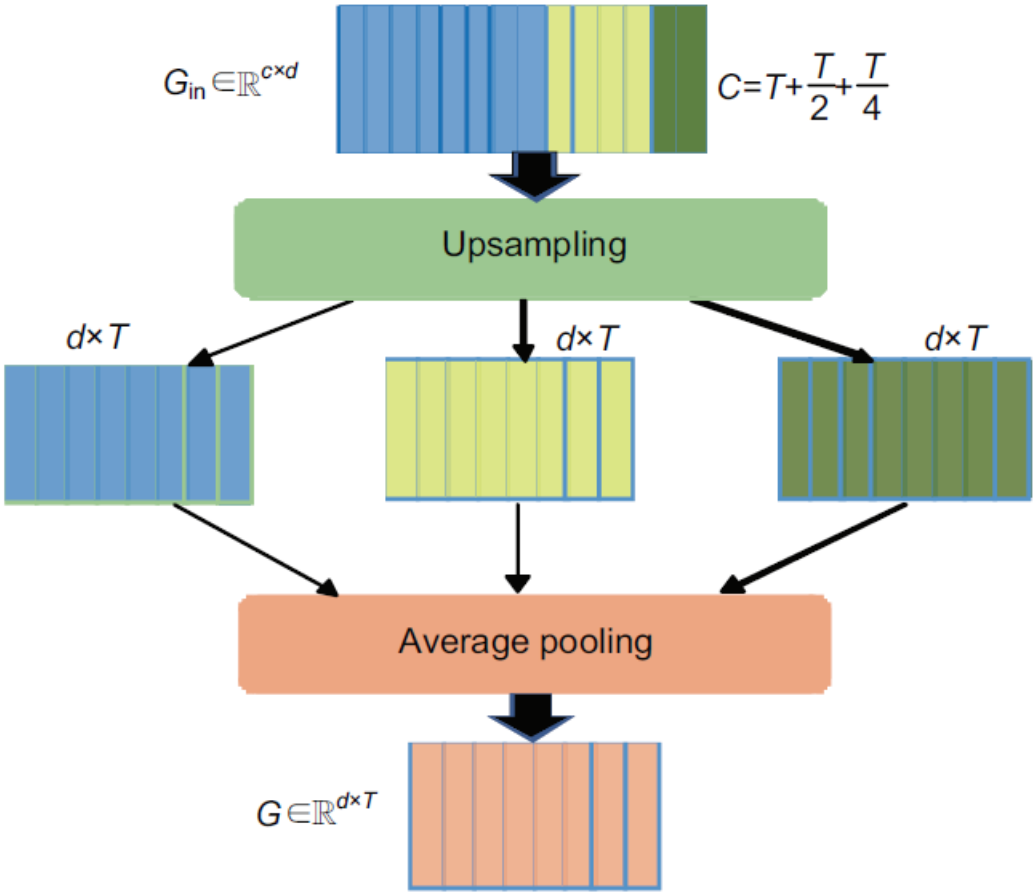
2. GCAN allows capturing a graph-structured context-aware representation of frames in a video, which facilitates identifying key frames with high importance scores.

# Major results




Framework of the graph convolutional attention network (GCAN)

# Major results (Cont'd)




Decoder of the graph branch in GCAN

# Major results (Cont'd)

Performance comparison with several supervised methods

| Method | F-score | | | | | |
|---|---|---|---|---|---|---|
| | SumMe | | | TVSum | | |
| | C | A | T | C | A | T |
| vsLSTM | 37.6 | 41.6 | 40.7 | 54.2 | 57.9 | 56.9 |
| dppLSTM | 38.6 | 42.9 | 41.8 | 54.7 | 59.6 | <u>58.7</u> |
| SUM-GAN$_{sup}$ | 41.7 | 43.6 | | 56.3 | 61.2 | |
| DR-DSN$_{sup}$ | 42.1 | 43.9 | 42.6 | 58.1 | 59.8 | **58.9** |
| HSA-RNN | | 44.1 | | | 59.8 | |
| DySeqDPP | 44.3 | | | 58.4 | | |
| SASUM$_{sup}$ | 45.3 | | | 58.2 | | |
| SUM-FCN | 47.5 | <u>51.1</u> | <u>44.1</u> | 56.8 | 59.2 | 58.2 |
| UnpairedVSN$_{psup}$ | 48.0 | | | 56.1 | | |
| CSNet$_{sup}$ | <u>48.6</u> | 48.7 | <u>44.1</u> | 58.5 | 57.1 | 57.4 |
| A-AVS | 43.9 | 44.6 | | 59.4 | 60.8 | |
| M-AVS | 44.4 | 46.1 | | **61.0** | **61.8** | |
| PCDL$_{sup}$ | 43.7 | 44.1 | | 59.2 | <u>61.3</u> | |
| GCAN$_{sup}$ | **53.0** | **54.2** | **46.8** | <u>60.7</u> | 61.1 | <u>58.7</u> |

The best records in each column are highlighted in bold, and the second-best ones are underlined explicitly. C: canonical; A: augmented; T: transfer

# Major results (Cont'd)

Performance comparison with individual components of $GCAN_{sup}$

| Parameter | F-score | | | | | |
|---|---|---|---|---|---|---|
| | SumMe | | | TVSum | | |
| | C | A | T | C | A | T |
| $CSNet_{sup}$ | 48.6 | 48.7 | 44.1 | 58.5 | 57.1 | 57.4 |
| $GCAN_{sup}$ | 53.0 | 54.2 | 46.8 | 60.7 | 61.1 | 58.7 |
| $GCAN_{temp}$ | 51.1 | 51.3 | 46.1 | 58.2 | 58.8 | 58.1 |
| $GCAN_{graph}$ | 51.5 | 50.3 | 44.6 | 59.3 | 59.5 | 57.9 |

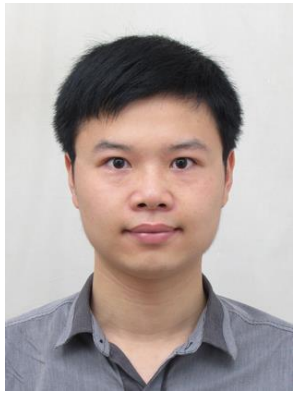
C: canonical; A: augmented; T: transfer

# Conclusions

1. This paper proposes a video summarization approach based on embedding learning and context fusion, named the graph convolutional attention network (GCAN).

2. GCAN consists of temporal graph embedding learning and context fusion, which derives an informative embedding that respects the intrinsic structure of video frames.

3. Empirical studies on two benchmarks demonstrate that GCAN enjoys better video summarization results.



Ping LI is an associate professor at Hangzhou Dianzi University. He received his PhD degree in computer science from Zhejiang University, China, in 2014. From 2016 to 2017, he was a post-doctor at National University of Singapore. His research interests include machine learning, computer vision, and multimedia computing.

Prof. Xianghua XU received his PhD degree from College of Computer Science and Technology, Zhejiang University, China. His research interests focus on data mining, Internet of Things, and artificial intelligence.