Tree-Augmented Cross-Modal Encoding for Complex-Query Video Retrieval

Xun Yang[1], **Jianfeng Dong**[2], Yixin Cao[1],
Xun Wang[2], Meng Wang[3], Tat-Seng Chua[1]

[1] National University of Singapore
[2] Zhejiang Gongshang University
[3] Hefei University of Technology

Text-to-Video Retrieval

- Give a textual query, the task is asked to retrieve semantically relevant videos from a list of candidate videos.
- How to represent textual queries matters.

Textual Query

Videos

Textual Queries in Video Retrieval

- From keyword based queries to more complex natural language sentence based queries.

Keyword based queries	puppy, play
Natural language sentence based queries	Two girls are laughing together and then another throws her folded laundry around the room

Related Work

Textual Query Representations	Papers
Word2vec+NetVLAD	Wray et al. ICCV19, Liu et al. BMVC19
Word2vec+mean pooling	Miech et al. ICCV19
Word2vec+FisherVector	Shao et al. ECCV18
LSTM/bi-LSTM	Yu et al. ECCV18, Yu et al. CVPR17
GRU/bi-GRU	Mithun et al. ICMR18,
<Subject, Verb, Object>+RNN	Xu et al. AAAI15
Multi-level (BoW, word2vec, GRU)	Dong et al. TMM18, Li et al. MM19
Multi-level (Local, Global, Temporal)	Dong et al. CVPR19
Graph Convolutional Networks	Chen et al. CVPR20

Related Work

GitHub https://github.com/danieljf24/awesome-video-text-retrieval

Awesome Video-Text Retrieval by Deep Learning awesome

Table of Contents

2020

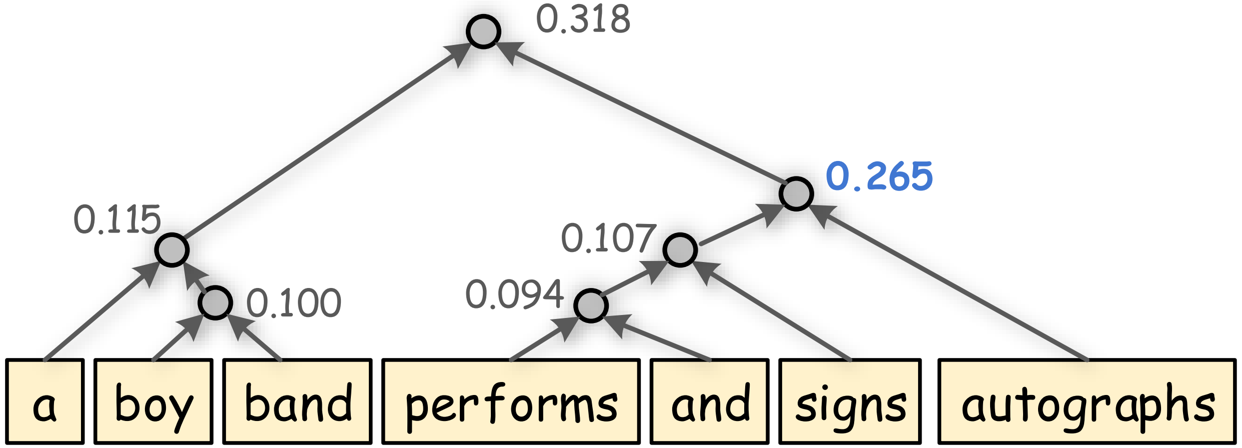
- `[Yang et al. SIGIR20]` Tree-Augmented Cross-Modal Encoding for Complex-Query Video Retrieval. SIGIR, 2020. [paper]
- `[Doughty et al. CVPR20]` Action Modifiers: Learning from Adverbs in Instructional Videos. CVPR, 2020. [paper]
- `[Chen et al. CVPR20]` Fine-grained Video-Text Retrieval with Hierarchical Graph Reasoning. CVPR, 2020. [paper]
- `[CVPR2020]`
- `[Zhu et al. CVPR20]` ActBERT: Learning Global-Local Video-Text Representations. CVPR, 2020. [paper]

2019

- `[Dong et al. CVPR19]` Dual Encoding for Zero-Example Video Retrieval. CVPR, 2019. [paper] [code]
- `[Song et al. CVPR19]` Polysemous visual-semantic embedding for cross-modal retrieval. CVPR, 2019. [paper]
- `[Wray et al. ICCV19]` Fine-Grained Action Retrieval Through Multiple Parts-of-Speech Embeddings. ICCV, 2019. [paper]
- `[Xiong et al. ICCV19]` A Graph-Based Framework to Bridge Movies and Synopses. ICCV, 2019. [paper]
- `[Li et al. ACMMM19]` W2VV++ Fully Deep Learning for Ad-hoc Video Search. ACM Multimedia, 2019. [paper] [code]
- `[Liu et al. BMVC19]` Use What You Have: Video Retrieval Using Representations From Collaborative Experts. MBVC, 2019. [paper] [code]
- `[Choi et al. BigMM19]` From Intra-Modal to Inter-Modal Space: Multi-Task Learning of Shared Representations for Cross-Modal Retrieval. International Conference on Multimedia Big Data, 2019. [paper]

Tree-augmented Query Encoder

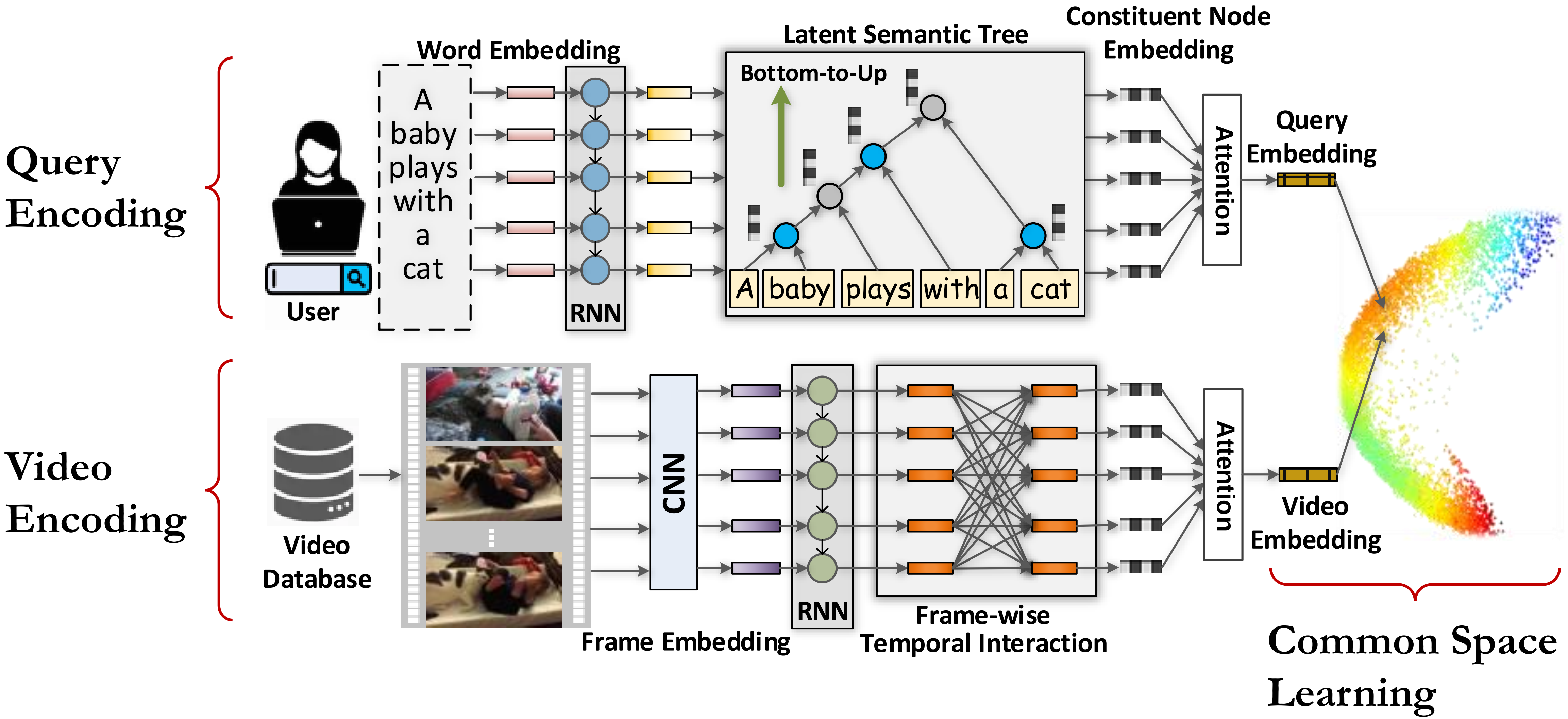
Query 1: a boy band performs and signs autographs



The tree is learned with the retrieval model in an end-to-end manner, **without any syntactic rules and annotations.**

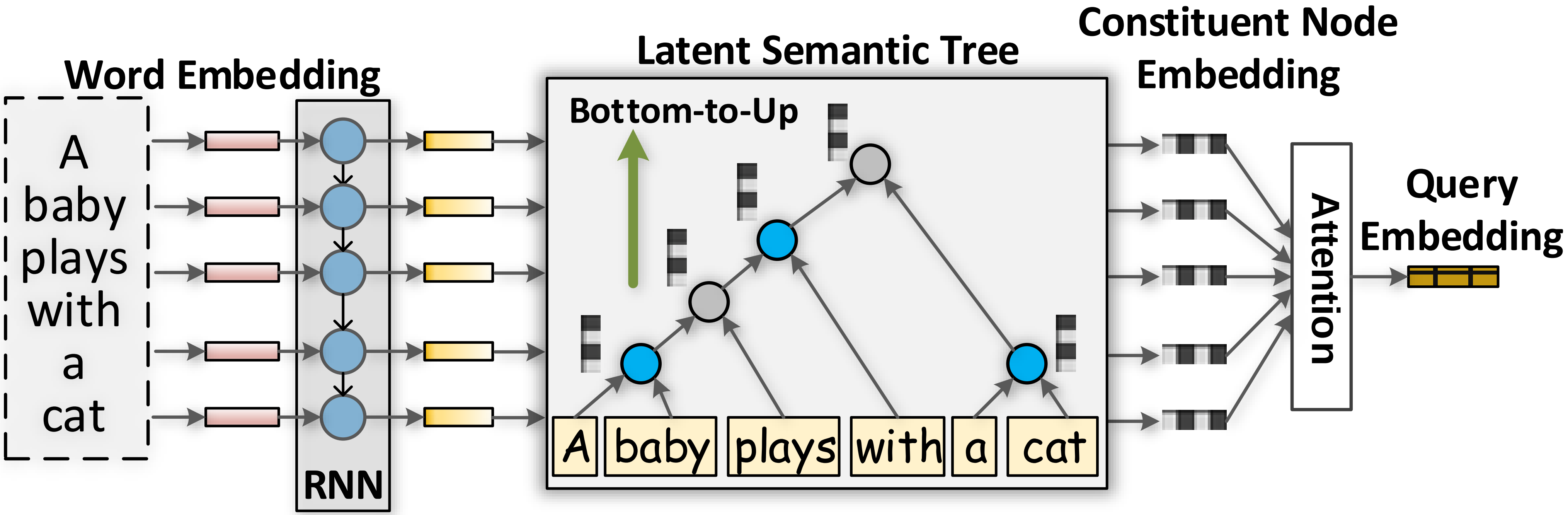
Our Method

Framework


Query Encoding
Word Embedding
Latent Semantic Tree
Constituent Node Embedding
Bottom-to-Up
A
baby
plays
with
a
cat
User
RNN
A baby plays with a cat
Attention
Query Embedding
Video Encoding
Video Database
CNN
RNN
Frame Embedding
Frame-wise Temporal Interaction
Attention
Video Embedding
Common Space Learning

Tree-augmented Query Encoder

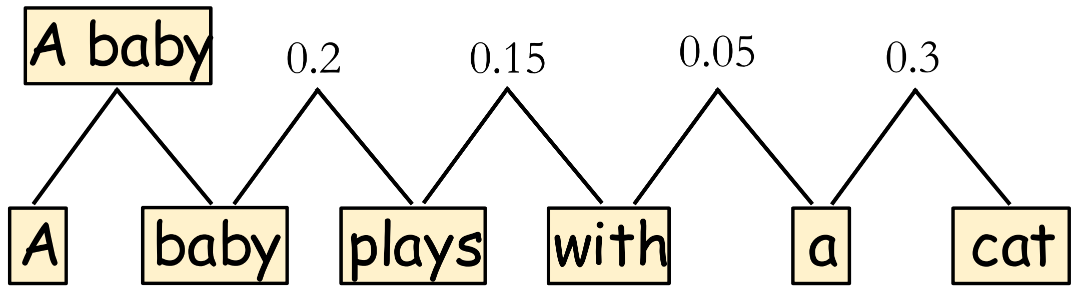
- We utilize Tree-structured LSTM (TreeLSTM) [Kai et al. ACL15] to recursively compose a latent semantic tree (LST) in a bottom-to-up fashion to structurally describe textual queries.

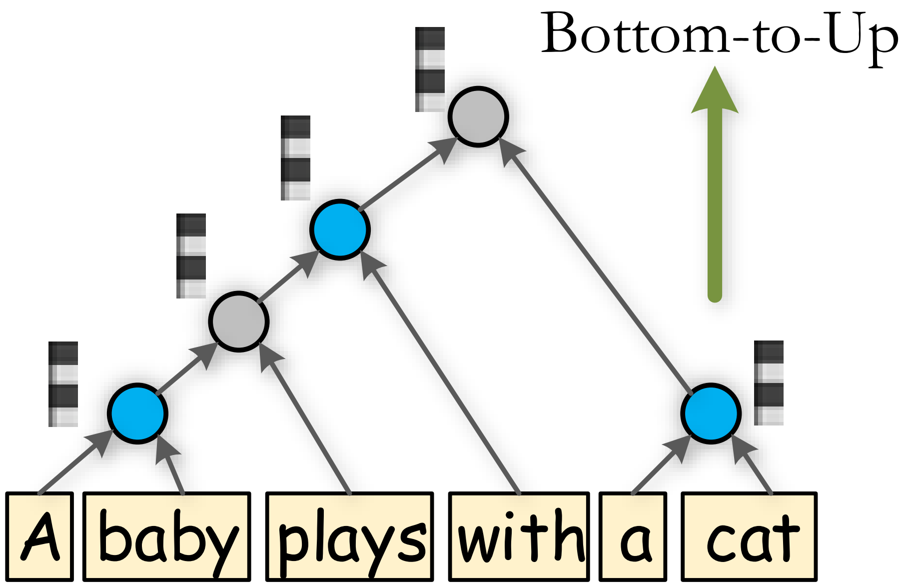


Kai et al. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In *ACL* 2015.

Latent Semantic Tree

- Select two adjacent child nodes to merge as a parent node
- Candidate parent node with the maximum score is regareder as the final true parent node
- Recursively repeat until only a single node is left





Node Representation

- Given the representations of two adjacent child nodes $(\mathbf{h}_i, \mathbf{c}_i)$ and $(\mathbf{h}_{i+1}, \mathbf{c}_{i+1})$ we use **TreeLSTM** to compute the **parent node representation.**

Left child node

Two forget gates

$$\begin{bmatrix} \mathbf{i} \\ \mathbf{f}_l \\ \mathbf{f}_r \\ \mathbf{o} \\ \mathbf{g} \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \left(\mathbf{W}^p \begin{bmatrix} \mathbf{h}_i \\ \mathbf{h}_{i+1} \end{bmatrix} + \mathbf{b}^p \right),$$

Right child node

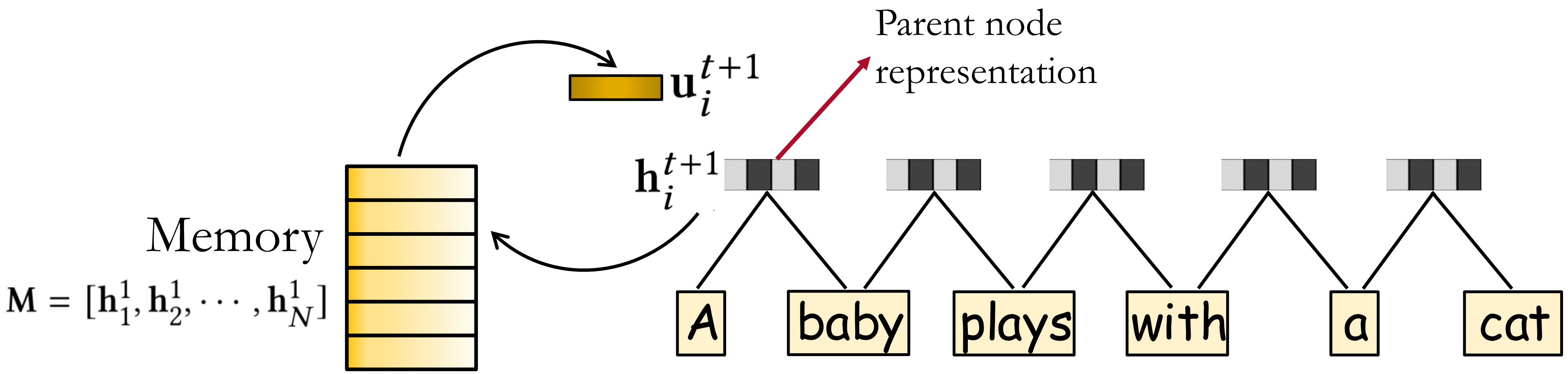
$$\mathbf{c}_p = \mathbf{f}_l \odot \mathbf{c}_i + \mathbf{f}_r \odot \mathbf{c}_{i+1} + \mathbf{i} \odot \mathbf{g},$$

Parent node representation

$$\mathbf{h}_p = \mathbf{o} \odot \tanh(\mathbf{c}_p),$$

Memory-augmented Node Scoring

- We propose a **memory-augmented node scoring and selection** to select two adjacent child nodes to merge as a parent node.



Attentive context feature

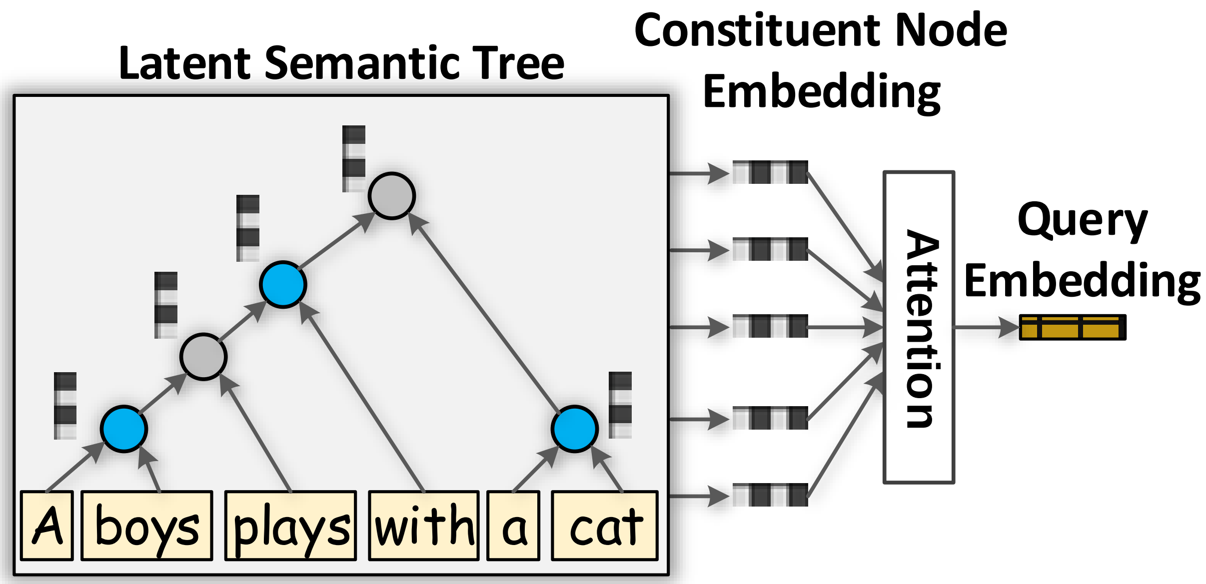
$$\mathbf{u}_i^{t+1} = (\mathbf{a}_i^{t+1})^{\mathrm{T}}\mathbf{M}, \quad a_{ij}^{t+1} = \mathrm{Softmax}\left((\mathbf{h}_i^{t+1})^{\mathrm{T}}\sigma(\mathbf{W}_m\mathbf{h}_j^1 + \mathbf{b}_m)/\sqrt{d_t}\right)$$

Node score

$$s_i^{t+1} = \mathrm{Softmax}\left(\mathbf{w}_s^{\mathrm{T}}\sigma\left(\mathbf{W}_s\begin{bmatrix}\mathbf{h}_i^{t+1}\\ \mathbf{u}_i^{t+1}\end{bmatrix} + \mathbf{b}_s\right)/\sqrt{2d_t}\right)$$

Structure-aware Query Representation

- We introduce an attention network to investigate the importance of each constituent and then derive the intention-aware query representation.



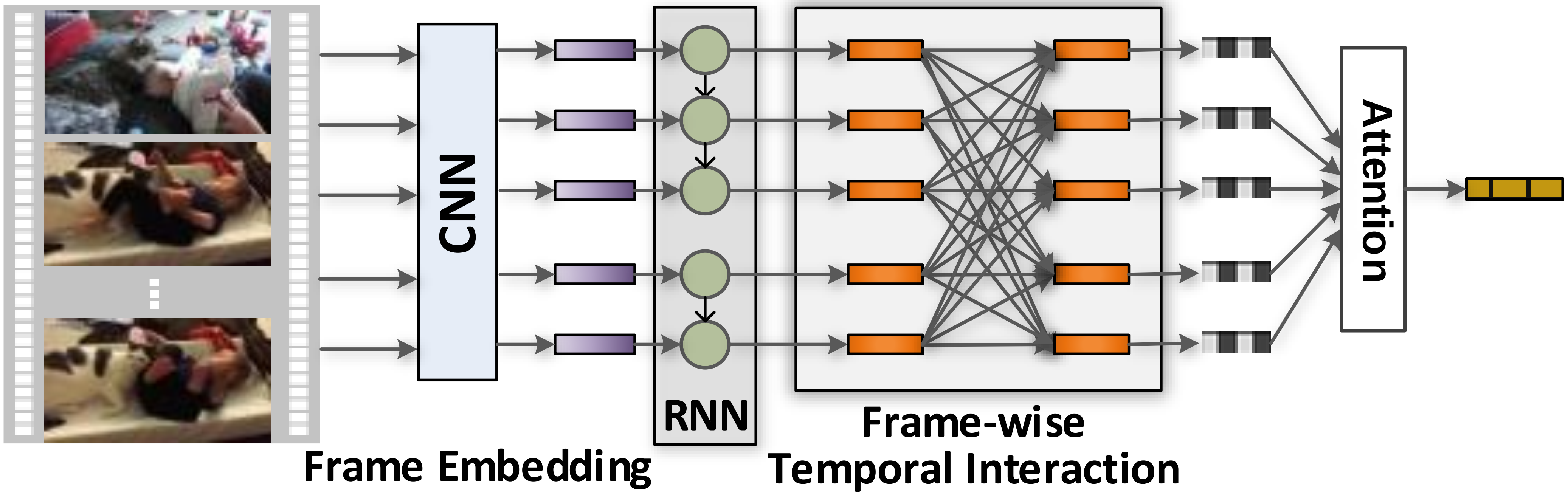
$$\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_{N-1}\} = \mathrm{LSTree}(\{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_N\})$$

$$\beta_i = \mathrm{Softmax}\left(\mathbf{u}_{ta}^{\mathrm{T}}\sigma(\mathbf{W}_{ta}\mathbf{e}_i + \mathbf{b}_{ta})/\sqrt{d_{ta}}\right)$$

$$\left.\bar{\mathbf{q}} = \sum_{i=1}^{N-1} \beta_i \mathbf{e}_i \right\} \text{Query representation}$$

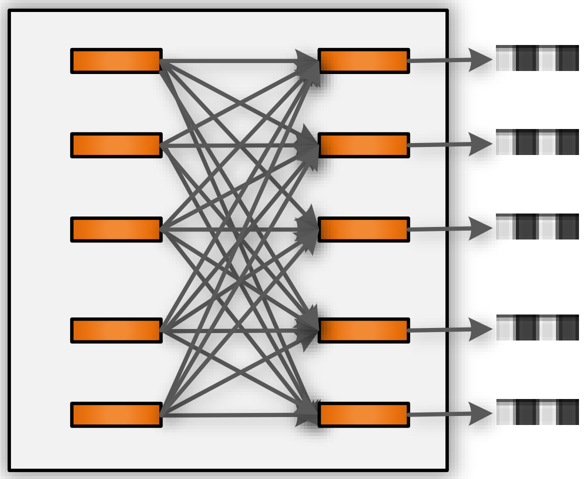
Temporal-Attentive Video Encoder

- We deal with two types of video characteristics: 1) **temporal dependence** between consecutive frames along the sequence and **frame-wise temporal interaction** over the whole video space.

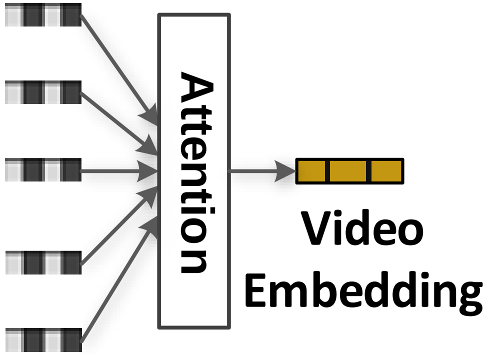


Temporal-Attentive Video Encoder

- To further enhance the representation of the video sequence, we propose to leverage the frame-wise correlation based on the **multi-head self-attention mechanism** [Ashish et al. NeuaIPS17].



Frame-wise Temporal Interaction



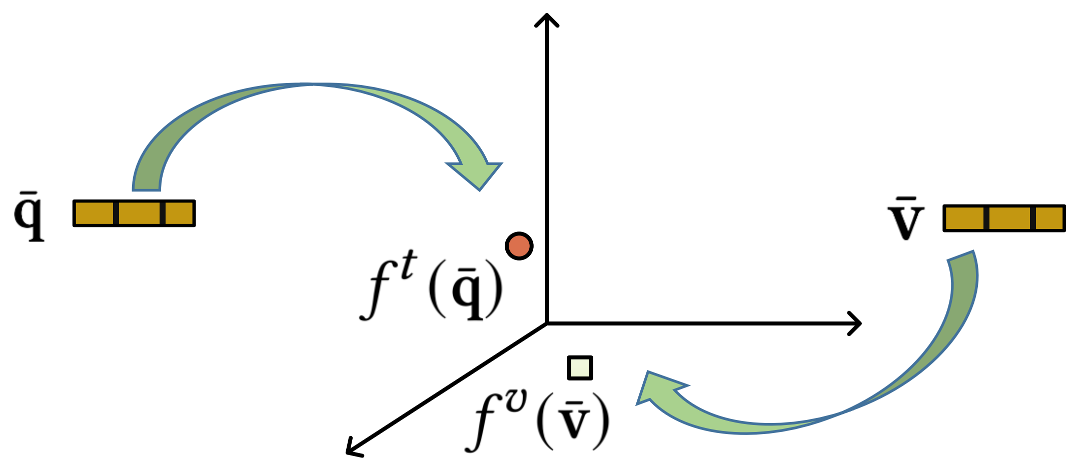
$$\eta_t = \text{Softmax}\left(\mathbf{u}_{va}^{\mathrm{T}} \sigma(\mathbf{W}_{va}\hat{\mathbf{v}}_t + \mathbf{b}_{va}) / \sqrt{d_{va}}\right)$$

$$\left.\bar{\mathbf{v}} = \sum_{t=1}^{M} \eta_t \hat{\mathbf{v}}_t \quad \right\} \text{Video representation}$$

Ashish et al. Attention is all you need. In *NeurIPS*, 2017.

Common Space Learning

- Given a textual query representation and a video representation, we project them into a common space by two linear projection matrices.



Triplet ranking loss with the hard negative mining [Faghri et al. BMVC2018]:

$$L(\mathcal{X}) = \frac{1}{|\mathcal{N}^h|} \sum_{i=1}^{B} \sum_{j \in \mathcal{N}^h} \max\left(0, \delta + s\left(\mathcal{Q}_i, \mathcal{V}_j\right) - s\left(\mathcal{Q}_i, \mathcal{V}_i\right)\right)$$

We just take into consideration the top $|\mathcal{N}^h|$ negative samples (e.g., 5) and average the costs for stable and efficient training.

Faghri et al. Improving Visual-Semantic Embeddings with Hard Negatives. In BMVC 2018.

Experiments

Experiments

- R1: How does the proposed method perform compared with state-of-the-art methods?
- R2: How the effects of the different components in our method?
- R3: How does the proposed method perform on different types of complex queries (e.g., different lengths)?

Performance comparison on MSR-VTT

- Our proposed TCE model consistently performs the best on three different data splits of MSR-VTT.

Method	**R@1**	**R@5**	**R@10**	**MedR**
Data split from [43]				
Dong *et al.* [6]	1.8	7.0	10.9	193
Mithun *et al.* [29]	5.8	17.6	25.2	61
DualEncoding [7]	7.7	22.0	31.8	32
TCE	**7.7**	**22.5**	**32.1**	**30**
Data split from [27]				
Random	0.3	0.7	1.1	502
CCA [42]	7.0	14.4	18.7	100
MEE [27]	12.9	36.4	51.8	10.0
MMEN (Caption) [42]	13.8	36.7	50.7	10.3
JPoSE [42]	14.3	38.1	53.0	**9**
TCE	**17.1**	**39.9**	**53.7**	**9**

Data split from [48]				
Random	0.1	0.5	1.0	500
C+LSTM+SA+FC7 [39]	4.2	12.9	19.9	55
VSE-LSTM [15]	3.8	12.7	17.1	66
SNUVL [49]	3.5	15.9	23.8	44
Kaufman *et al.* [14]	4.7	16.6	24.1	41
CT-SAN [50]	4.4	16.6	22.3	35
JSFusion [48]	10.2	31.2	43.2	13
Miech *et al.* [28]	12.1	35.0	48.0	12
TCE	**16.1**	**38.0**	**51.5**	**10**

Performance comparison on LSMDC

Method	R@1	R@5	R@10	MedR
C+LSTM+SA+FC7 [39]	4.3	12.6	18.9	98
VSE-LSTM [15]	3.1	10.4	16.5	79
SNUVL [49]	3.6	14.7	23.9	50
Kaufman *et al.* [14]	4.7	15.9	23.4	64
CT-SAN [50]	5.1	16.3	25.2	46
Miech *et al.* [26]	7.3	19.2	27.1	52
CCA (FV HGLMM) [16]	7.5	21.7	31.0	33
JSFusion [48]	9.1	21.2	34.1	36
Miech *et al.* . [28]	7.2	18.3	25.0	44
MEE [27]	10.2	25.0	33.1	**29**
TCE (Visual)	7.9	20.8	27.8	46
TCE (Visual+Mot.)	9.7	23.3	34.8	32
TCE (Visual+Mot.+Aud.)	**10.6**	**25.8**	**35.1**	**29**

- Our TCE again performs the best on LSMDC.
- TCE has the potential of improving its performance by leveraging more features

Experiments

- R1: How does the proposed method perform compared with state-of-the-art methods?
- R2: How the effects of the different components in our method?
- R3: How does the proposed method perform on different types of complex queries (e.g., different lengths)?

Ablation Studies on MSR-VTT

- Removing each component from TCE would result in relative performance degeneration, which shows the importance of each component.

Method	R@1	R@5	R@10	MedR
On Query Encoder				
WordEmb+AvgP	6.79	20.98	30.68	32
WordEmb+MaxP	5.92	18.90	27.82	40
LSTM	6.91	21.31	31.17	31
LSTM+AvgP	6.95	21.28	30.68	35
TCE (w/o-Cxt)	6.98	21.46	31.49	30
TCE (w/o-LSTM)	7.09	21.86	31.67	31
TCE (w/o-TAtt)+AvgP	6.59	20.57	30.48	34
TCE	**7.16**	**21.96**	**32.04**	30

→ Remove the memory

→ Remove the LSTM before LST

→ Remove attention and use mean pooling

Ablation Studies on MSR-VTT

- On video encoder, each component is also beneficial.

Method	R@1	R@5	R@10	MedR
On Video Encoder				
Frame+AvgP	6.67	20.41	29.89	36
Frame+MaxP	6.20	20.24	29.87	35
GRU	6.75	21.03	30.91	31
GRU+AvgP	6.17	19.51	28.71	38
TCE (w/o-Mha)	6.97	21.59	31.19	31
TCE (w/o-GRU)	7.08	21.96	31.86	30
TCE (w/o-VAtt)+AvgP	6.73	21.38	31.74	**29**
TCE	**7.16**	**21.96**	**32.04**	30

Remove the multi-head attention

Remove the GRU before LST

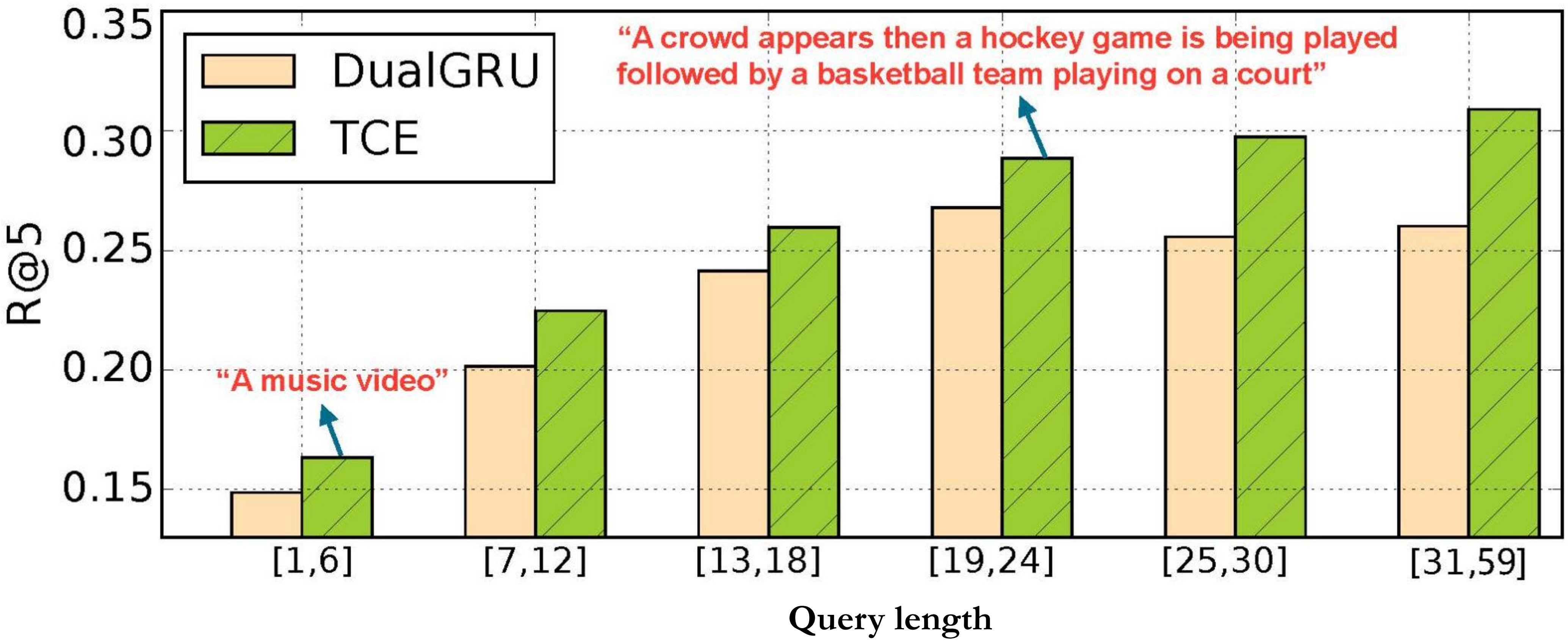
Remove attention and use mean pooling

Experiments

- R1: How does the proposed method perform compared with state-of-the-art methods?
- R2: How the effects of the different components in our method?
- R3: How does the proposed method perform on different types of complex queries (e.g., different lengths)?

Analysis on Different Types of Queries

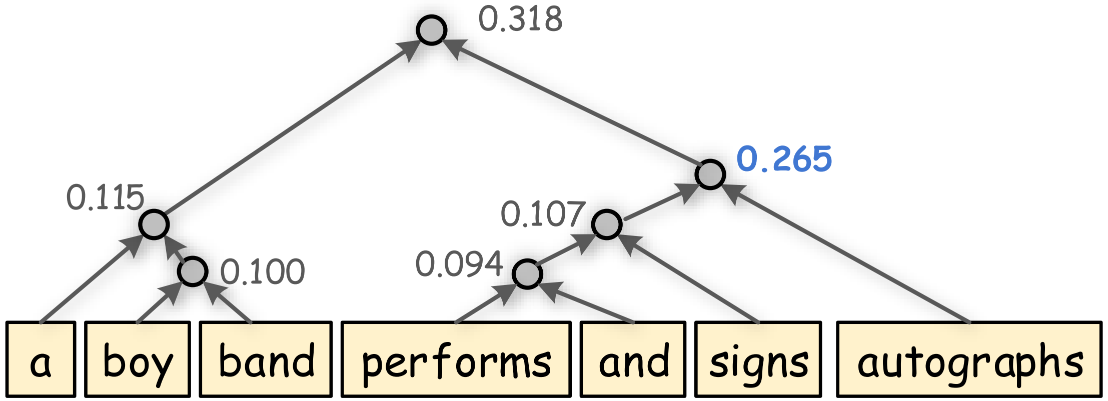
- Our proposed TCE is better to handle the complex queries.

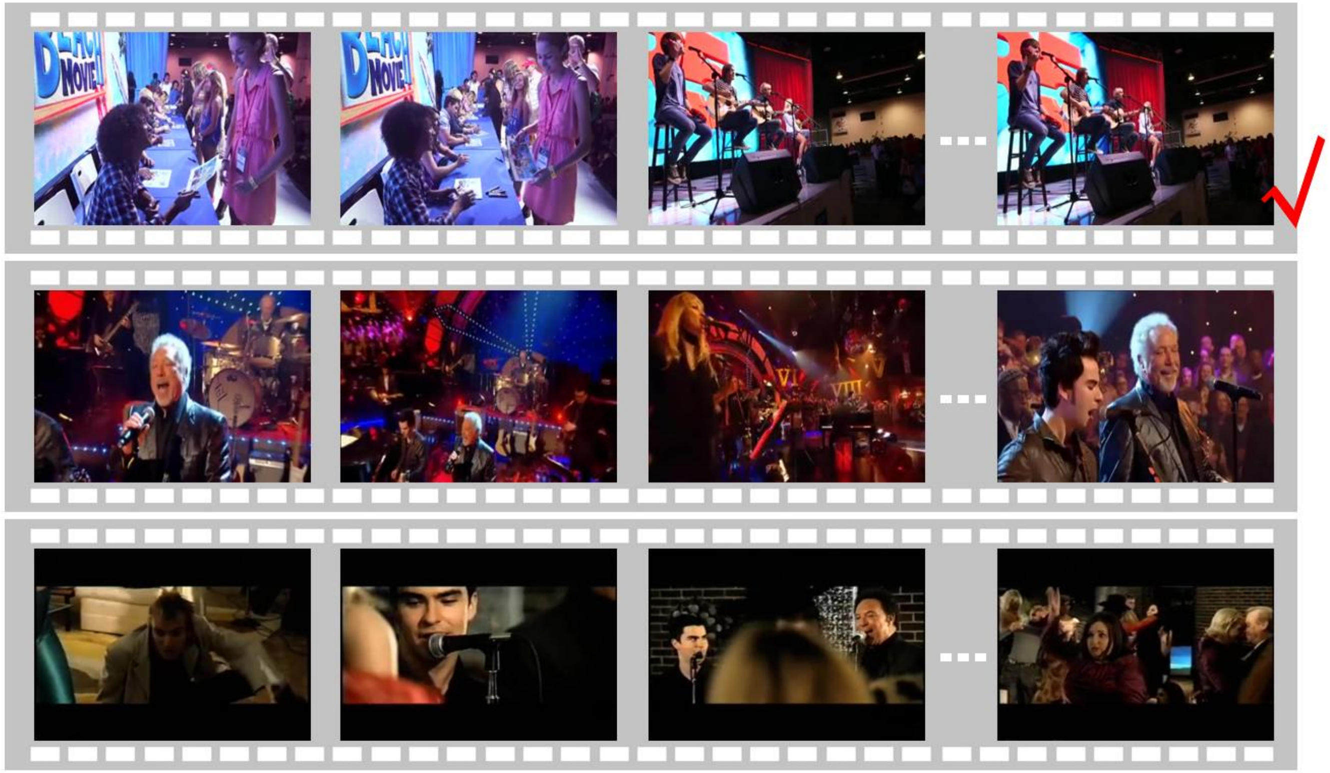


Qualitative Analysis

- Our proposed is able to construct syntactically reasonable tree.

Query 1: a boy band performs and signs autographs





Conclusions

- In this work, we proposed a novel method TCE for complex-query video retrieval, which consists of a **tree-based query encoder** and a **temporal attentive video encoder**. Extensive experiments on MSR-VTT and LSMDC datasets demonstrate its effectiveness.
- In the future, we will explore the proposed approach for other language-guided video tasks, such as video moment retrieval with natural language.
- We are also interested in exploring the **external knowledge** to enhance the text representation learning and the tree construction in the future study.

E-mail: xunyang@nus.edu.sg dongjf24@gmail.com