Time-Evolving Conditional Character-centric Graphs for Movie Understanding

Anonymous Author(s)
Affiliation
Address
email

Abstract

Temporal graph structure learning for long-term human-centric video understanding is promising but remains challenging due to the scarcity of dense graph annotations for long videos. It is the desired capability to learn the dynamic spatio-temporal interactions of human actors and other objects implicitly from visual information itself. Toward this goal, we present a novel Time-Evolving Conditional cHaracter-centric graph (TECH) for long-term human-centric video understanding with application in Movie QA. TECH is inherently a recurrent system of the query-conditioned dynamic graph that evolves over time along the story and follows throughout the course of a movie clip. As aiming toward human-centric video understanding, TECH uses a two-stage feature refinement process to draw attention to human characters and their interactions while treating the interactions with non-human objects as contextual information. Tested on the large-scale TVQA dataset, TECH clearly shows advantages over recent state-of-the-art models.

1 Introduction

Capturing the dynamic story in long-term human-centric video presents a powerful testbed for temporal graph modeling. An important setting is Movie QA, where the questions are to "probe" a certain aspect of the story – answering them would demonstrate a high degree of understanding. Here the challenges lie in the analysis of, and reasoning about the long-term temporal dynamic relationships of the human characters and the surroundings with the guidance of the question being asked. As movies are long and contain an abundant amount of information, it is also key to selectively attend to the most relevant visual entities in relation to the query. Current methods [4, 5, 7, 6, 10, 11, 12, 17] fall short of meeting these challenges.

Dynamic graph learning offers a natural scheme to solve these problems since graphs are coherent structures to represent the human-centric relationship within a movie clip. Toward this end, we propose a novel **T**ime-**E**volving **C**onditional c**H**aracter-centric graph (TECH) that inherits key advanced properties of temporal graph learning for long-term human-centric video understanding with application in Movie QA task. Unlike other approaches where graphs often stay static with pre-existed relations, TECH is a recurrent system of query-conditioned dynamic graphs that evolves along the time dimension. These dynamic structures allow TECH to effectively model the progressive nature of information in movie data. In addition, TECH introduces a two-stage representation refinement process that allows it to focus on the interactions between human characters given the contextual information of their interactions with the surrounding environment. Arguably, understanding the relationships across human characters is key in human-centric video understanding as these relationships are often the topic of interest by human queries. Tested on the large-scale TVQA dataset, we clearly demonstrate the modeling benefits of TECH against the most recent state-of-the-art models in Movie QA.

Submitted to NeurIPS 2022 Temporal Graph Learning Workshop. Do not distribute.

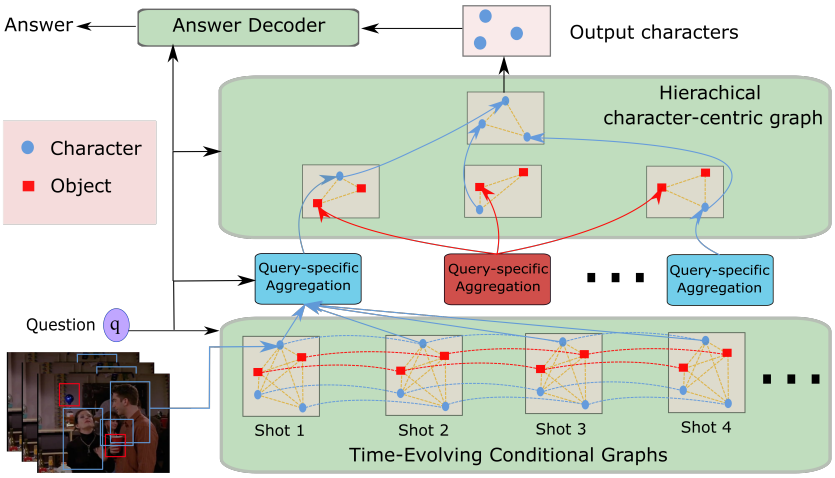

Figure 1: The overall structure of our proposed TECH for the Movie QA. Taking as input visual entities lives, TECH recurrently refines each entity representation with its surrounding entities living in the same shot and its own past state. The refined features are then passed through a query-specific information aggregation module to summarize the information carried by each entity over time. Later, TECH draws its main focus on the interactions between human characters using a hierarchical character-centric graph. Finally, an answer decoder is used to map the joint representation of multimodal inputs to the correct answer.

2 TECH: Time-Evolving Conditional Character-centric Graph

We present a novel method for solving Movie QA by representing the visual content of the movie as a *dynamic query-conditioned graph* of character-object that evolves over time.

Following [2], we detect and track visual features of entities (human characters and non-human objects) in a video, assuming that they live throughout the course of the video of S shots. This results in N connected sequences of entities over shots $V = \{z_{n,s}\}_{n=1,s=1}^{N,S}$. Details of visual processing are provided in the supplementary. The query and answer are encoded similar to those in [4, 11]: Each query is represented as contextual embedding matrix $Q = \{Q_p \in \mathbb{R}^{1\times d}\}_{p=1}^{L_Q}$, and a global vector $q_g = \text{mean_pooling}(\text{Q}) \in \mathbb{R}^d$.

Give these visual and linguistic features, we build our model termed **T**ime-**E**volving **C**onditional c**H**aracter-centric graph (TECH) to capture the dynamic interactions between human characters and their surroundings in response to the query. Our TECH is composed of three major modules: (1) Time-Evolving Conditional Graphs over Video Shots, (2) Query-specific Temporal Information Aggregation and (3) Hierarchical Character-centric Graph. The readout of TECH is a joint representation ready for answer prediction. Fig. 1 illustrates how these components interact with each other.

2.1 Time-Evolving Conditional Graphs over Video Shots

Shot-based Entities Graph: We build a dynamic graph $\mathcal{G}_s\left(V_s, E_s\right)$ for each video shot s based on the interactions between entities living in the same shot. Vertices V_s are a set of N entities $V_s = \{z_n(s)\}_{n=1}^{N}$taken from the representative frame of the shot s. For the sake of readability, we adhere refer to z_n as the feature of object n at shot s. The edges $E_s \in \mathbb{R}^{N\times N}$ are an adjacency matrix implying the relationships between the entities. We obtain E_s as a query-induced correlation matrix of the vertices' features:

$$E_s = \text{norm}\left(AA^\top\right); \text{ for } A = \{a_n\}_{n=1}^{N}, \text{and } a_n = \text{ReLU}\left(W_a\left([z_n, z_n \odot q_g]\right)\right), \tag{1}$$

where $norm$ is a normalization operator which is the softmax function in our implementation. W_a is learnable parameters.

Entity-based Graphs Evolving over Time: Recall that a video typically contains multiple video shots where each shot s is represented as the above defined graph $\mathcal{G}_s\left(V_s, E_s\right)$. It is crucial to effectively connect these shot-based graphs to reflect the continuous flow of the spatio-temporal nature of movie clips. In this work, we design a recurrent graph network that iteratively refines the representation of each vertex at the present shot s by the information received from its neighbors within the current shot and the information of itself at the previous shots. Mathematically, assuming $H_s^0 = \{z_n(s) \in \mathbb{R}^{1\times d}\}_{n=1}^N$ as the representation of the vertices at iteration 0, the refined representations of the vertices at iteration l are given by:

$$\begin{aligned} \text{TGCN}_s\left(H_s^{l-1}\right) &= W_2^{l-1}\text{ReLU}\left(E_s H_s^{l-1} W_1^{l-1} + b^{l-1} + H_{s-1}^L W_3^{l-1}\right), \\ H_s^l &= \text{ReLU}\left(H_s^{l-1} + \text{TGCN}_s\left(H_s^{l-1}\right)\right), \end{aligned} \tag{2}$$

where $l \in [1, L]$, E_s is the adjacency matrix given by Eq. 1. H_{s-1}^L is outputs of the refined representations of entities at the previous video shot. At the end of this stage, we have a set of S spatio-temporal representations for each of human characters and non-human objects across video shots $\bar{Z}_n = \left\{H_{n,s}^L \mid H_{n,s}^L \in \mathbb{R}^{1\times d}\right\}_{s=1}^S$.

2.2 Query-specific Temporal Information Aggregation

While visual content in a given movie clip may contain a large amount of information, the information relevant to a query is more specific. Thus, we design a *query-specific temporal information aggregation* module to retrieve visual moments that are only relevant to the query:

$$\tilde{z}_n = \frac{1}{L_Q}\sum_{p=1}^{L_Q}\text{Attention}(Q_p, \bar{Z}_n, \bar{Z}_n) \in \mathbb{R}^d, \tag{3}$$

where Q_p is the p-th word in the contextual query Q of length L_Q, and the attention function is defined in Eq. 5.

2.3 Hierarchical Character-centric Graph

In *human-centric* video understanding, there are two types of entity-level relations of interest: character-character and character-object. While one would naturally be more interested in the relationship between characters and how it develops throughout the course of a movie clip, the interactions between the characters with their surroundings often provide contextual information. Inspired by this observation, we design a *hierarchical character-centric graph* to characterize the different levels of interest in the character-object relationships as respect to human queries. In particular, we treat the visual entities $\tilde{Z} = \{\tilde{z}_n\}_{n=1}^N$ obtained by the aggregation function described above as of two different types: human characters $C = \{c_i\}_{i=1}^I \in \mathbb{R}^{\mathrm{I}\times\mathrm{d}}$, and non-human objects $O = \{o_j\}_{j=1}^P \in \mathbb{R}^{\mathrm{P}\times\mathrm{d}}$, where $\tilde{Z} = C \cup O$. The graph module operates through two stages of feature refinement: *object-to-character*, and *character-to-character*.

Object-to-Character Refinement For each detected human character c_i in C, we build a dynamic graph $\mathcal{G}_{c_i}$, whose nodes are the human character c_i and all other detected non-human objects in O. Let $X_i = \{c_i, O\}$ as the features of the nodes, we follow the procedure of building dynamic graphs with shot-based features in Sec. 2.1 to build the graph $\mathcal{G}_{c_i}$of node features X_i and an adjacency matrix A_i^{O2C}. Finally, we refine the representation of the nodes using Deep Graph Convolutional Network (DGCN) [9]: $\bar{X}_i = \text{DGCN}(X_i, A_i^{\mathrm{O2C}})$.

Character-to-Character Refinement Let $\bar{C} = \{\bar{x}_i\}_{i=1}^I$ be the the character embedding after the the last step. We then build a dynamic graph among only the characters with node features $\bar{C}$ and the relationships between the characters are denoted by an adjacency matrix A^{C2C}. We then again refine the representation of the characters with DGCN: $\hat{C} = \text{DGCN}(\bar{C}, A^{\mathrm{C2C}})$. Finally, we retrieve a joint representation, ready for answer decoding: $r = \frac{1}{L_Q}\sum_{p=1}^{L_Q}\text{Attention}(Q_p, \hat{C}, \hat{C}) \in \mathbb{R}^d$. The answer decoder is presented in the supplemental material.

3 Experiments

We evaluate the effectiveness of TECH on TVQA dataset [10]. As our main focus is to understand characters and the relationships with their surroundings solely from a visual perspective, we *do not use subtitles and timestamp annotations* in our experiments. The dataset contains a total of $21,763$ clips from six American TV series. Each movie clip is associated with seven multi-choice-related questions with five candidate answers, resulting in $152K$ questions.

Comparison against state-of-the-art methods Table 1 presents the performance of our proposed method TECH against most recent state-of-the-art (SOTA) methods on the TVQA validation set. For fair comparisons, we do not mention models that are extensively pretrained on large-scale video-text data or models that use external dense captions such as [1, 3, 4, 18, 19]. As showed, TECH sets a new state-of-the-art result by achieving 47.79% in QA accuracy on the validation set, which is around 2 absolute points better than the latest advances in Movie QA. The results confirm the benefits of utilizing structured data by TECH compared to existing methods in movie understanding task.

Models	Val. Acc↑
TVQA w. CNN feat. [10]	42.01
TVQA w. visual concept [10]	44.27
BERT Video QA [16]	44.63
STAGE [11]	45.83
DenseCap* [4]	45.85
TECH	**47.79**

Table 1: Performance of TECH in comparison with state-of-the-art methods on TVQA dataset (* = no captions used).

Ablation studies To provide more insights of TECH, we conduct multiple ablation studies to examine the contributions of individual modules in TECH. Details of the experimental results are given in Table 2.

Model	Val. Acc.↑
Full model	47.79
w/o query-induced mat.	46.95
w/o time-evolving graph	46.97
w/o query-specific agg.	47.30
w/o char. centric graph	47.10

Table 2: Ablation studies on the TVQA dataset

▷ *Effectiveness of the query-induced correlation matrix*: We examine the contribution of our query-guided correlation matrix as described in Sec. 2.1. In particular, we deliberately set the values of the edges of the graphs with uniform values. This results in a considerable decrease in performance, confirming that TECH effectively selects meaningful connections between characters and their surrounding objects toward answering a given query.

▷ *Effectiveness of the time-evolving conditional graphs*: To verify the impact of our time-evolving conditional graphs, we ablate the recurrent graph refinement step as in Sec. 2.1 from the full model. This degrades the performance of TECH by around 0.8 points. The result clearly demonstrates the significance of modeling the evolving of information over time in movie understanding.

▷ *Effectiveness of the query-specific temporal information aggregation*: Rather than using the aggregation operator as in Sec. 2.2, we utilize the mean pooling operation instead. Results show that doing this would lead to a marginal decrease in performance.

▷ *Effectiveness of hierarchical character-centric graph*: We investigate the effectiveness of our hierarchical character-centric graph by replacing the structure as in Sec. 2.3 with the simple average pooling operator to get the representation as an input of the answer decoder. Results suggest that this would hurt the model's performance by nearly 0.7 points.

4 Conclusion

This paper introduced a **T**ime-**E**volving **C**onditional c**H**aracter-centric graph (TECH) for long-term human-centric video understanding with application in Movie QA. Built upon sequences of human characters and non-human objects living in space-time, we designed TECH as a recurrent system of query-dependent dynamic graphs that allow information to effectively flow from early points to later points in time. The design of TECH has shown that modeling temporal graphs was crucial for movie understanding as it helped reflect the evolution of events in movie clips. TECH also showed the benefits of paying attention to human characters and their interactions within a movie clip over the interactions with other non-human objects. With advantages in model design, TECH clearly advances recent state-of-the-art Movie QA models on the large-scale TVQA dataset.

References

[1] Aman Chadha, Gurneet Arora, and Navpreet Kaloty. iperceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering. *arXiv preprint arXiv:2011.07735*, 2020. 3

[2] Long Hoang Dang, Thao Minh Le, Vuong Le, and Truyen Tran. Hierarchical object-oriented spatio-temporal reasoning for video question answering. *IJCAI*, 2021. 2, A.1

[3] Aisha Urooj Khan, Amir Mazaheri, Niels Da Vitoria Lobo, and Mubarak Shah. Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering. *arXiv preprint arXiv:2010.14095*, 2020. 3

[4] Hyounghun Kim, Zineng Tang, and Mohit Bansal. Dense-caption matching and frame-selection gating for temporal localization in videoqa. *arXiv preprint arXiv:2005.06409*, 2020. 1, 2, 3, 3

[5] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. Gaining extra supervision via multi-task learning for multi-modal video question answering. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 1

[6] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. Progressive attention memory network for movie story question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8337–8346, 2019. 1

[7] Junyeong Kim, Minuk Ma, Trung Pham, Kyungsu Kim, and Chang D Yoo. Modality shifting attention network for multi-modal video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10115, 2020. 1

[8] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. A.2

[9] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017. 2.3

[10] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. Tvqa: Localized, compositional video question answering, 2018. 1, 3

[11] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*, 2019. 1, 2, 3

[12] Fei Liu, Jing Liu, Xinxin Zhu, Richang Hong, and Hanqing Lu. *Dual Hierarchical Temporal Convolutional Network with QA-Aware Dynamic Normalization for Video Story Question Answering*, page 4253–4261. Association for Computing Machinery, New York, NY, USA, 2020. 1

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. A.2

[14] Tomáš Souček and Jakub Lokoč. Transnet v2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020. A.1

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. A.1

[16] Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. BERT Representations for Video Question Answering. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1556–1565, 2020. 3

[17] Zhaoquan Yuan, Siyuan Sun, Lixin Duan, Xiao Wu, and Changsheng Xu. Adversarial multi-modal network for movie question answering. *arXiv preprint arXiv:1906.09844*, 2019. 1

[18] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16375–16387, June 2022. 3

[19] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021. 3

A Supplemental Material

A.1 Model Details

Visual features A movie clip typically contains multiple video shots by nature. Hence, we divide each video input into shots using the common shot transition detection algorithm TransNet V2 [14]. For each video shot, we parse it into visual entities of human characters and non-human objects using object detection. These detected human characters and non-human objects are tracked and linked across shots using their visual salience and geometrical positions within respective video frames following [2]. Formally, given a video with S shots, we only take one key frame per shot to reduce computation cost. We now represent the given video as a total of N sequences of entities with appearance features denoted by $Z^a = \{z^a_{n,s} \mid z^a_{n,s} \in \mathbb{R}^{1024}\}^{N,S}_{n=1,s=1}$ and positional features denoted by $Z^p = \{z^p_{n,s} \mid z^p_{n,s} \in \mathbb{R}^4\}^{N,S}_{n=1,s=1}$, then obtain *position-aware visual features* of the detected entities at shot s using the following multiplicative gating mechanism:

$$z_{n,s} = \tanh(W_{z^a} z^a_{n,s} + b_{z^a}) \odot \sigma(W_{z^p} z^p_{n,s} + b_{z^p}), \tag{4}$$

where $\sigma(\cdot) \in (0, 1)$ is the sigmoid function and $\odot$ denotes element-wise product between two vectors.

Attention function We utilize the attention mechanism [15] that takes as input a triplet of query $q \in \mathbb{R}^{1\times d}$, keys $K \in \mathbb{R}^{M\times d}$, and values $V \in \mathbb{R}^{M\times d}$:

$$\text{Attention}(q, K, V) := \sum_{m=1}^{M} \text{softmax}_m(\frac{K_m W_k (qW_q)^\top}{\sqrt{d}}) V_m W_v \in \mathbb{R}^{1\times d}, \tag{5}$$

where W_k, W_q and W_v are learnable parameters.

Answer decoder As the task Movie QA in TVQA is defined as a multiple choice test, given input as the joint representation r and a set $Q_g = \{q_{g,i}\}^5_{i=1}$ represented five question-answer choice pairs, we use a MLP network with two fully-connected layers, then simply normalize the output by the softmax function to rank answer candidates:

$$y_i = \text{MLP}\left(W_r[r; W_{q_g} q_{g,i} + b_{q_g}] + b_r\right), \tag{6}$$

$$y^{\text{prob}}_i = \text{softmax}_i\left(W_y y_i + b_y\right). \tag{7}$$

Finally, the predicted answer is taken as $\bar{y} = \text{argmax}_i \left(\left\{ y^{\text{prob}}_i \right\}^5_{i=1} \right)$.

Loss function In this paper, the cross-entropy is selected as the loss function to train our model.

A.2 Implementation Details

We use Faster-RCNN trained on MS-COCO for frame-wise character-object detection and DeepSort for multiple object tracking. The RoBERTa pretrained model [13] is utilized to embed the QA pairs into vectors with a dimension size of 768. The number of characters and objects per video is 4 and 6, respectively. The dimension of hidden features in our model is set to 128. The default setting of GCN layers is 2. We apply the Adam optimizer [8] with an initial learning rate of 10^{-4}.

A.3 Qualitative Analysis

Figure 2 provides examples where TECH handles successfully while DenseCap struggles. Keys to answering these questions lie in the capability to capture the long-term temporal dynamics of human characters with their surrounding environment. DenseCap without explicit long-term temporal dependencies modeling struggles to predict correct answers.

Question: What did Monica do after she walked in the door ?

Answer candidates:

A. grabbed a bottle of water | B. set her purse down | C. took her phone out

D. showed Rachel a check | E. jumped up and down with joy

Ground truth: B. set her purse down | TECH: B. set her purse down | DenseCap: D. showed Rachel a check

Question: What color mug does Joey have next to him when Monica sits down next to him ?

Answer candidates:

A. Green | B. Red | C. Yellow | D. Blue | E. White

Ground truth: D. Blue | TECH: D. Blue | DenseCap: C. Yellow

Question: Where did House put his left hand after the doctors left ?

Answer candidates:

A. On his chest | B. Over her mouth | C. On Cuddy 's shoulder

D. On his waist | E. In his pocket

Ground truth: D. On his waist | TECH: D. On his waist | DenseCap: B. Over her mouth

Question: What does Paul do after Castle hands him the papers ?

Answer candidates:

A. Paul goes to the other room | B. Paul goes to sleep | C. Paul runs away

D. Paul starts reading | E. Paul faints

Ground truth: D. Paul starts reading | TECH: D. Paul starts reading | DenseCap: A. Paul goes to the other room

Figure 2: Qualitative examples show advantages of TECH in handling long-term temporal relationships in video while DenseCap struggles.