Video Graph Transformer for Video Question Answering

J. Xiao, P. Zhou, T. S. Chua, and S. Yan, "Video graph transformer for video question answering," in Proc. ECCV 2022, 2022, pp. 39-58.

Full Name : Ravialdy Hidayat

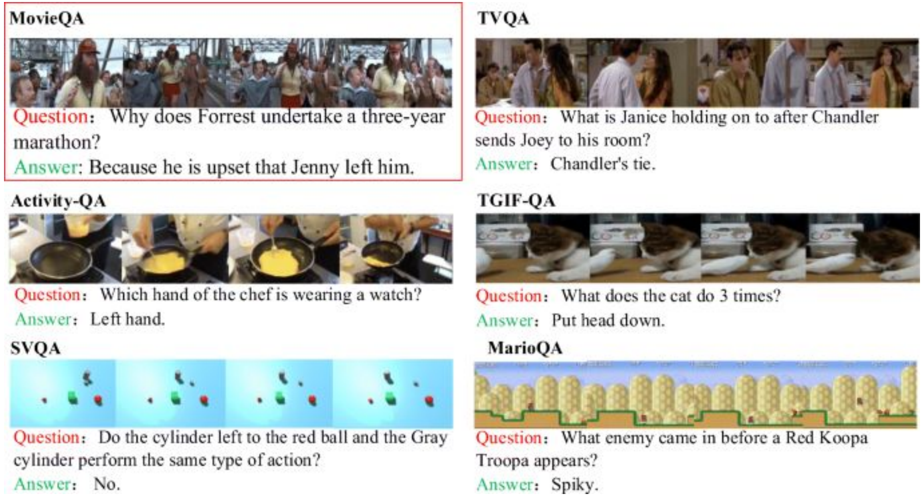
Student ID Number : 22110779

Advisor : Professor Cheol Jeong

VLI Lab

SEJONG UNIVERSITY

Background

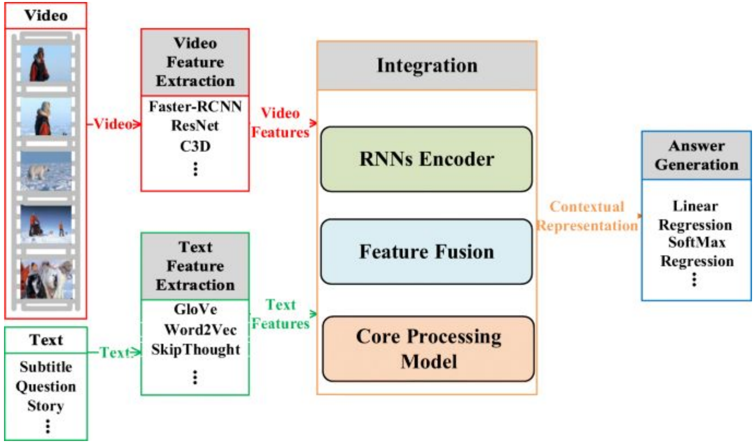


Source Image: Sun, G., Liang, L., Li, T. et al. Video Question Answering: a Survey of Models and Datasets. Mobile Netw Appl 26, 1904–1937 (2021). https://doi.org/10.1007/s11036-020-01730-0

- **Current SOTA** Video Question Answering (**VQA**) models **need large scale video-text data**.
- **Current Transformer based models** for **VideoQA** are oftenly considered as having **low performance on visual reasoning**. The **authors argue** that there are **two major reasons** why those problems occur.

Source: J. Xiao, P. Zhou, T. S. Chua, and S. Yan, "Video graph transformer for video question answering," in Proc. ECCV 2022, 2022, pp. 39-58.

#1 Major Reason : Video encoders are overly simplistic !!

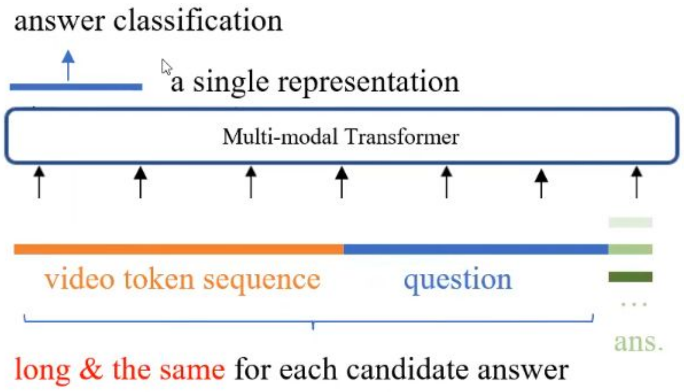


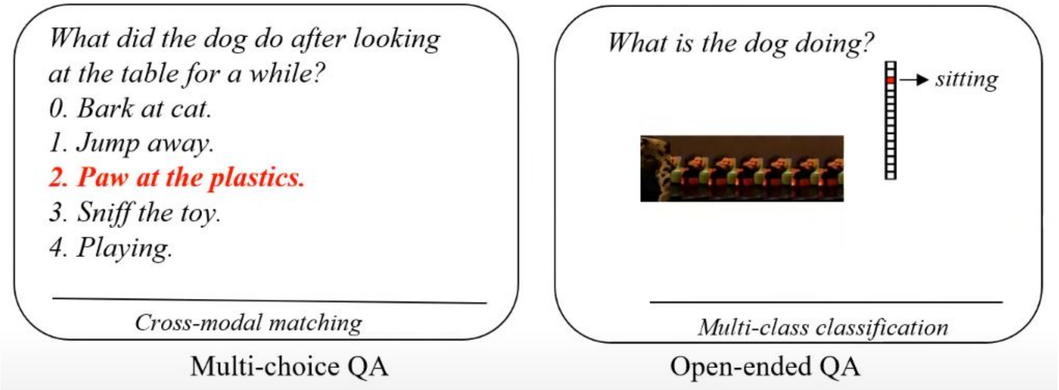
Source Image: Sun, G., Liang, L., Li, T. et al. Video Question Answering: a Survey of Models and Datasets. Mobile Netw Appl 26, 1904–1937 (2021). https://doi.org/10.1007/s11036-020-01730-0

- **Video encoders** used **nowadays** typically are **CNNs** or **Transformer implemented on 2D or 3D neural networks**, usually over short video segments.
- These approaches **can encode videos holistically**, **but** oftenly **fail** to **model spatio-temporal interactions** between **visual objects**.
- Therefore, those method's performances are **weak in visual relation reasoning** and **need large amount of data** to overcome that issue.

Source: J. Xiao, P. Zhou, T. S. Chua, and S. Yan, "Video graph transformer for video question answering," in Proc. ECCV 2022, 2022, pp. 39-58.

#2 Major Reason : Inappropriate Formulation of VideoQA Problem !





- In case of **multi-choice QA**, the **importance** of **short answers** will be **less** due to shorter representations compared with video and question ones, leading to a **weak generated global representation** in disambiguating the candidate answers.
- In the case of **open-ended QA**, it is common to formulate the problem as a multi-class classification problem where **answers are treated as class indexes**. **These** kind of **approaches ignore** the **role of their word semantics**.

Source: J. Xiao, P. Zhou, T. S. Chua, and S. Yan, "Video graph transformer for video question answering," in Proc. ECCV 2022, 2022, pp. 39-58.

Solution for #1 Major Problem

- Designs a **Dynamic Graph Transformer (DGT)** for video encoder that can **explicitly capture visual objects**, the **relations** of them, and their dynamics for **spatial and temporal** relation **reasoning**.
- The authors argue that **both** multi-choice and open-ended QA have an **objective to maximize this** following **function**.

$$a^* = \arg\max_{a \in \mathcal{A}} \mathcal{F}_W(a|q, v, \mathcal{A})$$

where, a^* is the final answer generated by the model

$\mathcal{A}$ is the set of candidate (multi-choice QA) or global answers (open-ended QA)

$\mathcal{F}_W$ is a mapping function with learnable parameters W

f^{qv} is a query-aware video representation

f^{a} is the candidate answer representation

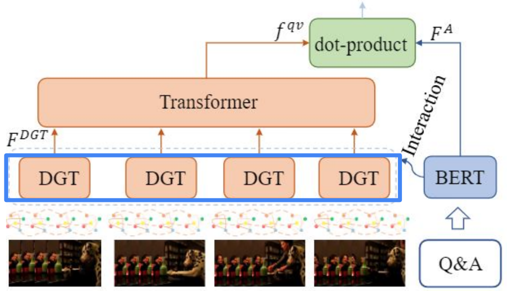
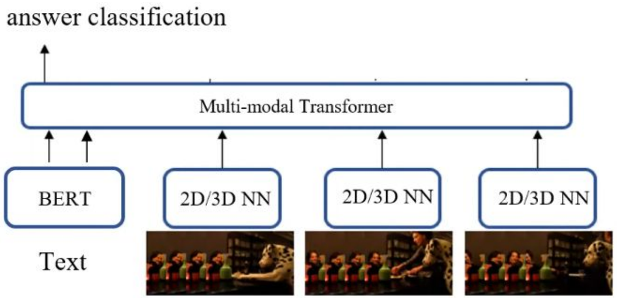


Fig. 1. Overview of video graph transformer (VGT) for VideoQA.

- Proposed model **Video Graph Transformer (VGT)** is created to perform $\mathcal{F}_W$ mapping function.
- This model will **represent the query-relevant video content** by integrating textual and visual object graphs information.
- The final answer will be generated by calculating similarity using dot-product between f^{qv} and f^{a}

Source: J. Xiao, P. Zhou, T. S. Chua, and S. Yan, "Video graph transformer for video question answering," in Proc. ECCV 2022, 2022, pp. 39-58.

Solution for #2 Major Problem



Typical Architecture

1st Transformer for Vision

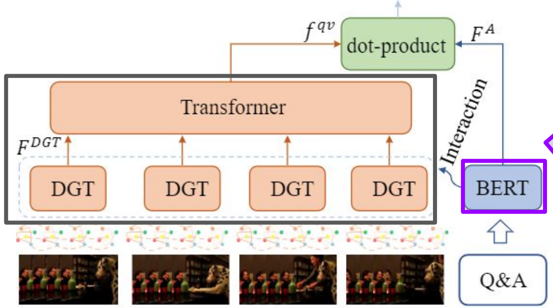


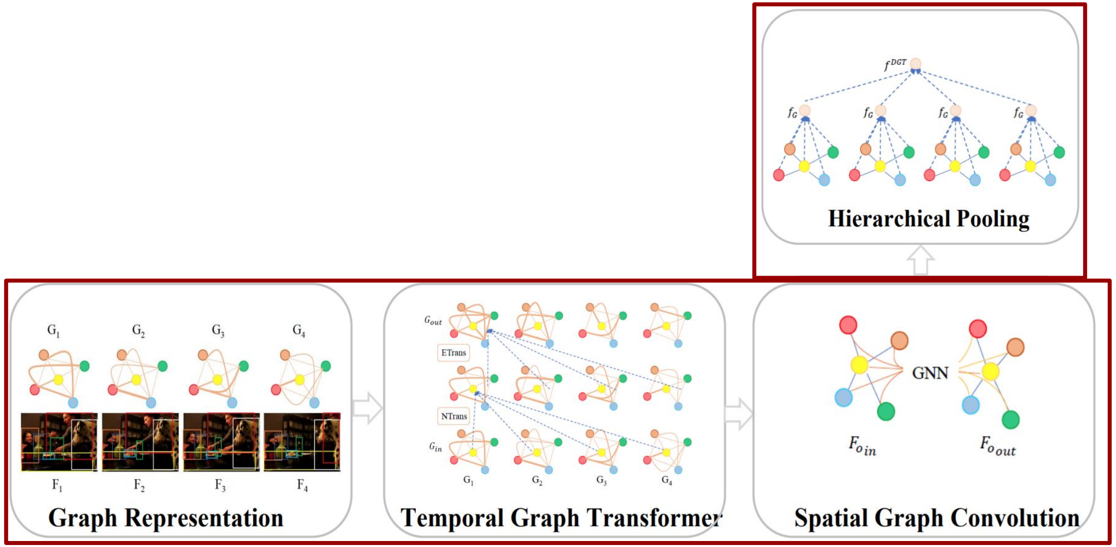
Fig. 1. Overview of video graph transformer (VGT) for VideoQA.

2nd Transformer for Text

- **VGT model separates vision and text transformers** to encode video and text respectively so that in the end it can calculate the similarity between them, while **common approach just use a single cross-modal transformer** to integrate extracted vision and text features.
- An additional module called **cross-modal interaction** is created to **communicate** between **vision** and **text** informations.

Source: J. Xiao, P. Zhou, T. S. Chua, and S. Yan, "Video graph transformer for video question answering," in Proc. ECCV 2022, 2022, pp. 39-58.

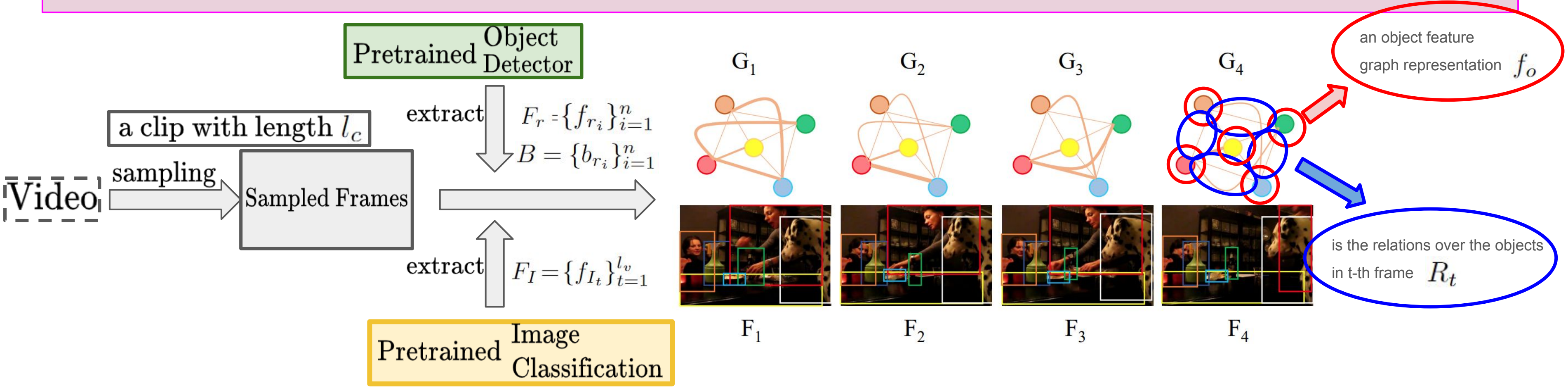
Dynamic Graph Transformer



- **Dynamic Graph Transformer** (**DGT**) **consists** of **four stages**, such as Graph Representation, Temporal Graph Transformer, Spatial Graph Convolution, and Hierarchical Pooling.
- It uses **contextual graphs** to improve the graphs representations obtained at static frames.
- It also **starts with local interactions**, **then go to global activities** to get the hierarchical view.

Source: J. Xiao, P. Zhou, T. S. Chua, and S. Yan, "Video graph transformer for video question answering," in Proc. ECCV 2022, 2022, pp. 39-58.

1st Stage : Video Graph Representation



$$f_o = \mathrm{ELU}(\phi_{W_o}([f_r; f_{loc}])) \Rightarrow F_o = \{f_{o_i}\}_{i=1}^{n} \Rightarrow R_t = \sigma(\phi_{W_{ak}}(F_{o_t})\phi_{W_{av}}(F_{o_t})^\top), \quad t \in \{1, 2, \ldots, l_v\} \Rightarrow \boxed{G_t = (F_{o_t}, R_t)}$$

where,

f_r is the object appearance representations generated by the object detector, and F_r is the sequence of f_r

f_{loc} is the location representations obtained by applying 1x1 convolution to spatial representations along time dimension.

l_v is the total sampled frames
k is the number of clips

F_{o_t} is the node representations of the graph in the t-th frame.

R_t is the edge representations of the graph in the t-th frame.

G_t is the graph representations of the graph in the t-th frame.

B is the sequence of spatial location representations of all objects in a frame $\{b_{r_i}\}_{i=1}^{n}$

l_c is the length of a clip that is defined as $\boxed{l_c = \frac{l_v}{k}}$

$\phi_{W_{ak}}$ is the linear transformations with parameters W_{ak}

ϕ_{W_o} is the linear transformations with parameters W_o

$\phi_{W_{av}}$ is the linear transformations with parameters W_{av}

F_I is the sequence of image-level feature from all sampled frames $\{f_{I_t}\}_{t=1}^{l_v}$

Source: J. Xiao, P. Zhou, T. S. Chua, and S. Yan, "Video graph transformer for video question answering," in Proc. ECCV 2022, 2022, pp. 39-58.

2^{nd} Stage : Temporal Graph Transformer

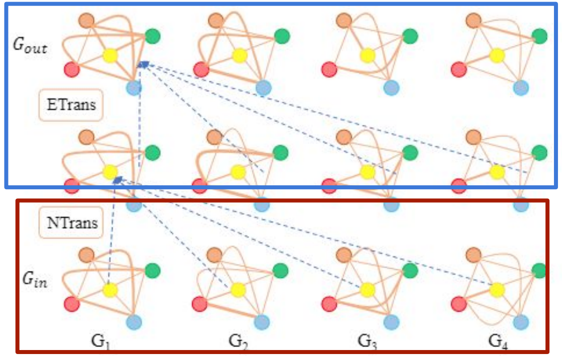


Fig. 3. Illustration of temporal graph transformer in a short video clip.

$$F'_{o_i} = \mathrm{NTrans}(F_{o_i}) = \mathrm{MHSA}^{(H)}(F_{o_i})$$

$$\mathcal{R} = \{R_t\}_{t=1}^{l} \in \mathbb{R}^{l_c \times d_n} \ (d_n = n^2)$$

$$\mathcal{R}' = \mathrm{ETrans}(\mathcal{R}) = \mathrm{MHSA}^{(H)}(\mathcal{R})$$

$$G_{out_t} = (F'_{o_t}, R'_t)$$

Notes :

$\mathrm{MHSA}^{(H)}$ is Multi-Head Self Attention

- **NTrans** will **aggregate object representations from other nodes** of the **same object** in all adjacent frames **within a clip**.

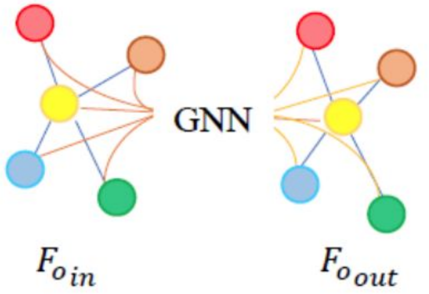
- **ETrans** is used to **model** the **temporal relation dynamics explicitly**.

- Temporal graph transformer takes a set of graphs G_{in} as an input and outputs a new set of graphs G_{out}.
- Use two main methods, such as **Node Transformer** (**NTrans**) and an **Edge Transformer** (**ETrans**).
- **NTrans** is created **to model the change of single object behaviours** so that it **can infer dynamic actions**.
- While, **ETrans** can **help** to **calibrate** the **wrong relations** and **recall** the **missing ones**.

Source: J. Xiao, P. Zhou, T. S. Chua, and S. Yan, "Video graph transformer for video question answering," in Proc. ECCV 2022, 2022, pp. 39-58.

3rd Stage : Spatial Graph Convolution

- Recall that **Temporal Graph Transformer focuses on temporal relation reasoning**, but the model **still needs the reasoning** capability **over the objects spatial interactions**.
- To do that, the **authors apply a U-layer graph attention convolution**.



$$F_o'^{(u)} = \mathrm{ReLU}((R' + I)F_o'^{(u-1)}W^{(u)}) \Rightarrow F_{o_{out}} = F_o' + F_o'^{(U)}$$

The last skip connection.

where,

$W^{(u)}$ is the graph parameters at the u-th layer.

I is the identity matrix for skip connections.

F_o' is the previous updated node representations, index t is omitted for brevity.

$F_o'^{(u)}$ is the output node representations of F_o', index t is omitted for brevity.

Source: J. Xiao, P. Zhou, T. S. Chua, and S. Yan, "Video graph transformer for video question answering," in Proc. ECCV 2022, 2022, pp. 39-58.

4^{th} Stage : About Hierarchical Aggregation

Sequence of clip-level feature

$F^{\mathrm{DGT}} = \{f_c^{\mathrm{DGT}}\}_{c=1}^{k}$

$f^{\mathrm{DGT}} = \mathrm{MPool}(F_G) = \frac{1}{l_c}\sum_{t=1}^{l_v} f_{G_t}$

3rd step : Then, aggregate all of the frame-level graph representations to obtain clip-level feature f^{DGT} .

2nd step : Concatenate with image representations since it can lose sight of global view of a frame.

$f_G = \mathrm{ELU}(\phi_{W_m}([\phi_{W_f}(f_I); f_G]))$

$f_G = \sum_{i=1}^{N} \alpha_i F_{o_{out_i}}, \quad \alpha = \sigma(\phi_{W_G}(F_{o_{out}}))$

1st step : Aggregate all objects graph representation in the same frame to obtain frame-level feature.

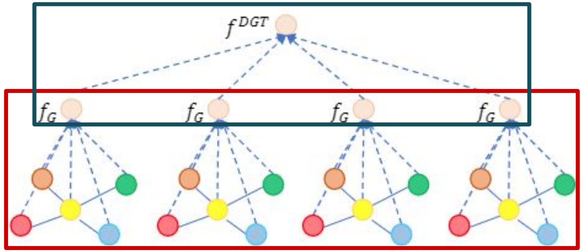


Fig. 4. Hierarchical Aggregation.

- The node representations have already recognized the objects' spatial and temporal interactions, but **those kind of interactions still lack of global perspective.**
- For aggregating these local interactions into higher-level video elements, the authors implement a hierarchical aggregation strategy like Fig.4.

where,

ϕ_{W_G} is the linear transformation with parameters $W_G \in \mathbb{R}^{d\times 1}$

$F_{o_{out}}$ is the output of previous spatial graph convolution.

f_G is the frame-level graph representation.

f_I is the frame-level image representation.

ϕ_{W_m} is the linear transformation with parameters $W_m \in \mathbb{R}^{2d\times d}$

ϕ_{W_f} is the linear transformation with parameters $W_f \in \mathbb{R}^{2048\times d}$

Source: J. Xiao, P. Zhou, T. S. Chua, and S. Yan, "Video graph transformer for video question answering," in Proc. ECCV 2022, 2022, pp. 39-58.

About Cross-Modal Interaction

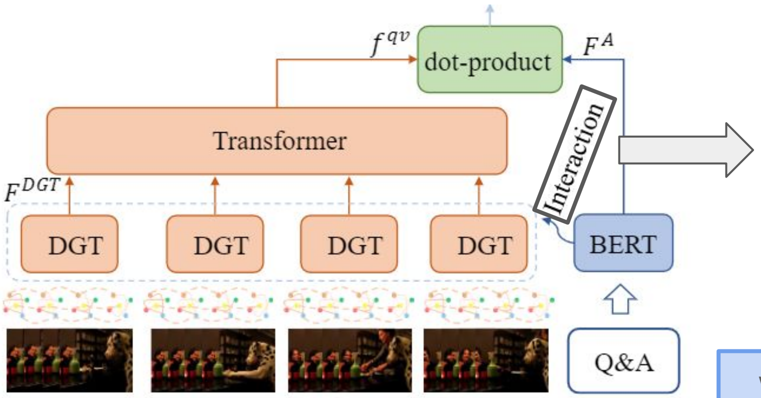
- **Information between vision and textual nodes** is done by using an additional module called **cross-modal attention**.
- For vision information, the input is a set of visual nodes X^v, while the input for textual information is a sequence

$X^q = \{x^q_m\}_{m=1}^{M}$ where *M* is the number of tokens in the text query.

Notes :

x^v: visual representations, e.g., F^{DGT}

x^q: textual representations, e.g., Outputs from BERT.



Cross-Modal Interaction

$$x^{qv} = x^v + \sum_{m=1}^{M} \beta_m x^q_m, \quad \text{where} \quad \beta = \sigma(x^v (X^q)^\top)$$

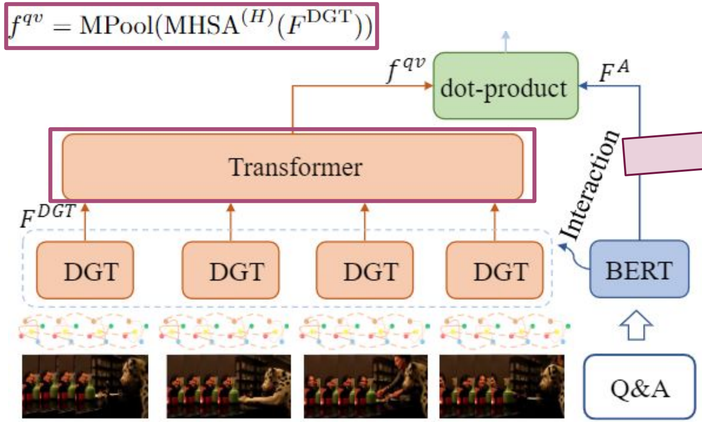
where,

ϕ_{W_Q} is the linear transformation with learnable parameters $W_Q \in \mathbb{R}^{768 \times d}$

x^{qv} is the video-query representations

Source: J. Xiao, P. Zhou, T. S. Chua, and S. Yan, "Video graph transformer for video question answering," in Proc. ECCV 2022, 2022, pp. 39-58.

About Global Transformer



- The created **DGT module** already pays an attention to **extract meaningful visual information from clips**.
- But we also **need an additional module** to **capture** informative **temporal dynamics** between **those clips**.
- The authors **implement another *H*-layer transformer** over the cross-modal interacted clip feature with **trainable positional embeddings**.

The authors argue that the implementation of **this global transformer has several advantages**, such as :

- **Retains the overall hierarchical structure** which can keep informative features from various level of the video elements.
- **Can further benefit the cross-modal interaction between vision and textual** informations since it can improve both features compatibility.

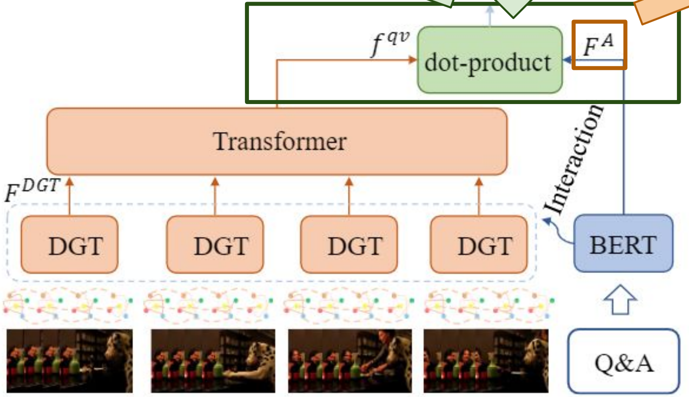
Source: J. Xiao, P. Zhou, T. S. Chua, and S. Yan, "Video graph transformer for video question answering," in Proc. ECCV 2022, 2022, pp. 39-58.

Answer Prediction

$$a^* = \arg\max(s)$$

$$s = f^{qv}(F^A)^\top \qquad s = f^{qv}(F^A)^\top \odot f^q(F^A)^\top$$

Multi-choice QA | Open-ended QA



$F^A = \{f_a^A\}_{a=1}^{|\mathcal{A}|}$, where $f^A = \mathrm{MPool}(X^A)$

where,

- X^A is the candidate answer's token representations using the help of BERT.
- $|\mathcal{A}|$ is the the number of candidate answers.
- f^q is the question representation generated by the same approach as
- $\odot$ is the element-wise product.
- a^* is the predicted answer's token representation.
- f^A is the candidate answer's token representation.

Source: J. Xiao, P. Zhou, T. S. Chua, and S. Yan, "Video graph transformer for video question answering," in Proc. ECCV 2022, 2022, pp. 39-58.

Experiment

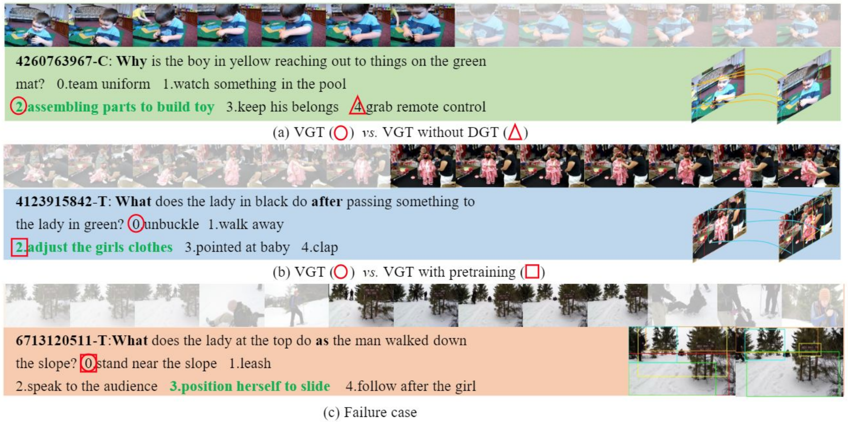


Fig. 7. Result visualization on NExT-QA [59]. The ground-truth answers are in green.

Models	CM-Pretrain	TGIF-QA					MSRVTT-QA
		Action	Transition	FrameQA	Action†	Transition†	
LGCN [19]	-	74.3	81.1	56.3	-	-	-
HGA [23]	-	75.4	81.0	55.1	-	-	35.5
HCRN [28]	-	75.0	81.4	55.9	55.7	63.9	35.6
B2A [41]	-	75.9	82.6	57.5	-	-	36.9
HOSTR [10]	-	75.0	83.0	58.0	-	-	35.9
HAIR [36]	-	77.8	82.3	60.2	-	-	36.9
MASN [47]	-	84.4	87.4	59.5	-	-	35.2
PGAT [42]	-	80.6	85.7	61.1	58.7	65.9	38.1
HQGA [60]	-	76.9	85.6	61.3	-	-	38.6
MHN [43]	-	83.5	90.8	58.1	-	-	38.6
ClipBERT [29]	VG+COCO Caption	82.8	87.8	60.3	-	-	37.4
SiaSRea [67]	VG+COCO Caption	79.7	85.3	60.2	-	-	41.6
MERLOT [70]	Youtube180M, CC3M	94.0	96.2	**69.5**	-	-	**43.1**
VGT (Ours)	-	**95.0**	**97.6**	61.6	**59.9**	**70.5**	39.7

Method	CM-Pretrain	NExT-QA Val				NExT-QA Test			
		Acc@C	Acc@T	Acc@D	Acc@All	Acc@C	Acc@T	Acc@D	Acc@All
HGA [23]	-	46.26	50.74	59.33	49.74	48.13	49.08	57.79	50.01
IGV [35]	-	-	-	-	-	48.56	51.67	59.64	51.34
HQGA [60]	-	48.48	51.24	61.65	51.42	49.04	**52.28**	59.43	51.75
P3D-G [9]	-	51.33	52.30	62.58	53.40	-	-	-	-
VQA-T* [64]	-	41.66	44.11	59.97	45.30	42.05	42.75	55.87	44.54
VQA-T* [64]	How2VQA69M	49.60	51.49	63.19	52.32	47.89	50.02	61.87	50.83
VGT (Ours)	-	**52.28**	**55.09**	**64.09**	**55.02**	**51.62**	51.94	**63.65**	**53.68**

- **VGT** successfully **outperforms previous SoTA** models on tasks that challenge **temporal dynamic reasoning significantly**.
- **VGT's performance** even **surpasses** those **methods** that are **pretrained** on **large-scale** vision-text **data**.

Source: J. Xiao, P. Zhou, T. S. Chua, and S. Yan, "Video graph transformer for video question answering," in Proc. ECCV 2022, 2022, pp. 39-58.

Conclusion

Video Graph Transformer for Video Question Answering

Junbin Xiao[1,2,3], Pan Zhou[1], Tat Seng Chua[2,3], and Shuicheng Yan[1]

[1] Sea AI Lab
[2] Sea-NExT Joint Lab, Singapore
[3] Department of Computer Science, National University of Singapore
junbin@comp.nus.edu.sg, zhoupan@sea.com, dcscts@nus.edu.sg, yansc@sea.com

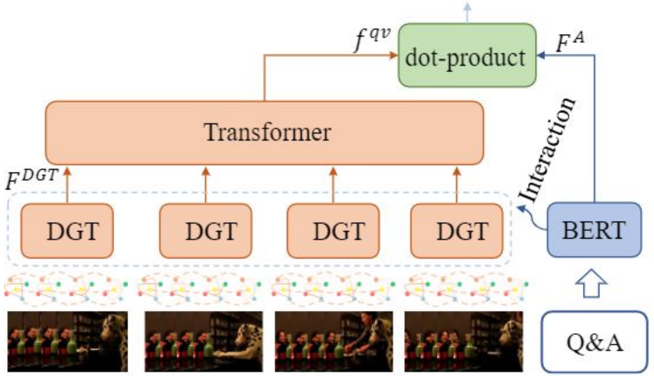


Fig. 1. Overview of video graph transformer (VGT) for VideoQA.

- The author propose a new model called **VGT or Video Graph Transformer** that has **two unique components,** Dynamic Graph Transformer (DGT) and separate transformer for vision and text informations with cross-modal interaction as the communication tool for both of them.
- **Its performances** successfully **surpass SOTA** models that are **pretrained** with **millions** of **external data**.

Source: J. Xiao, P. Zhou, T. S. Chua, and S. Yan, "Video graph transformer for video question answering," in Proc. ECCV 2022, 2022, pp. 39-58.

Thank you