Video Graph Transformer for Video Question Answering

Junbin Xiao, Pan Zhou, Tat-Seng Chua and Shuicheng Yan
Sea AI Lab (SAIL) & National University of Singapore (NUS)

Read on your phone

Watch our video

Data & Code

Introduction:

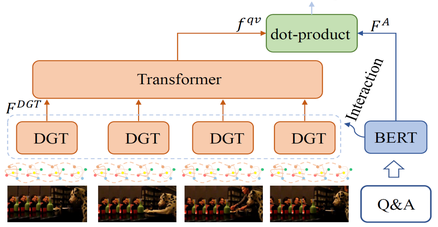
- Existing transformer-style models only demonstrate their success in answering questions that involve the coarse recognition or description of video contents. Their performances are either unknown or weak in answering questions that challenge real-world visual relation reasoning, especially the causal and temporal relations that feature video dynamics at action- and event-level. Cross-modal pretraining seems promising, yet it requires the handling of million-scale *video*-text data.

MSRVTT-QA & MSVD-QA [Xu et al, MM'17]:
Who is looking at the dog? Lady.
What is the dog doing? Sitting.

NExT-QA[Xiao et al, CVPR'21]:
Why did the woman walk towards the table in the middle of the video? Clean the table.

Method:

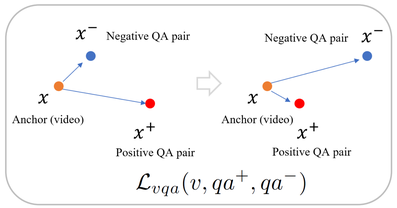
- We propose Video Graph Transformer (**VGT**) to improve previous arts in answering relation-type questions from two major aspects:
  - **Video Encoding:** we maintain a local-to-global hierarchical architecture and design dynamic graph transformer (DGT) that explicitly encodes the visual objects, their relations and dynamics, for spatial and temporal relation reasoning.
  - **Supervised Contrastive Learning**: we design *separated* video and text transformers to encode video and QA information respectively, for contrastive learning instead of cross-modal transformer for answer classification. Cross-modal interaction is done by additional light-weight cross-modal interaction



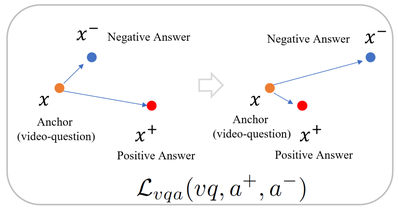
Global transformer to temporally localize the referring expression in questions, e.g., "*woman walk towards the table*"

DGT to perform fine-gained human-object interaction reasoning, and drive the answer, e.g., "*clean the table*"

- Contrastive Learning



(a) Multi-choice QA



(b) Open-ended QA

$$\mathcal{L}_*(x, x^+, x^-) = -\mathbb{E}_i[\log(\frac{e^{s_{\text{VGT}}(x_i, x_i^+)}}{e^{s_{\text{VGT}}(x_i, x_i^+)} + \sum_{(x_i, x_j^-)\in\mathcal{N}_i} e^{s_{\text{VGT}}(x_i, x_j^-)}})]$$

- Dynamic Graph Transformer
  - **Spatial-temporal**:
    Consider contextual graphs to improve the graphs obtained at static frames.
  - **Compositional**:
    Summarize local/atomic interactions to global activities.

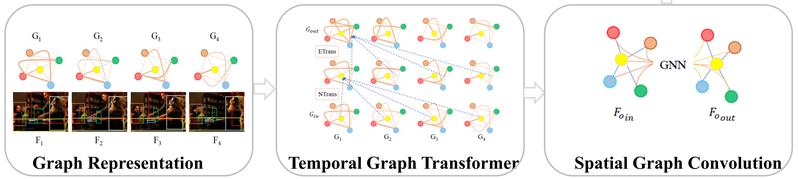


Illustration of the 4 stages to encode a video clip.

- Cross-modal Interaction

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

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

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

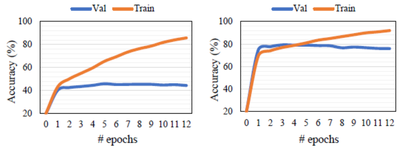
Experiment:

- SoTA Comparison.

Methods	NExT-Val	NExT-Test
VQA-T*[ICCV'21]	45.30	44.54
HQGA[AAAI'22]	51.42	51.75
VQA-T* (PT)	52.32	50.83
P3D-G[AAAI'22]	53.40	-
VGT (Ours)	55.02	53.68
VGT(PT)	**56.89**	**55.70**

Methods	TGIF-QA		TGIF-QA-R*	
	Action	Trans	Action	Trans
PGAT[MM'21]	80.6	85.7	58.7	65.9
ClipBERT[CVPR'21]	82.8	87.8	-	-
MERLOT[NeurIPS'21]	94.0	96.2	-	-
VGT (Ours)	**95.0**	**97.6**	59.9	70.5
VGT(PT)	-	-	**60.5**	**71.5**

Methods	TGIF-FQA	MSRVTT-QA
CoMVT[CVPR'21]	-	37.3
ClipBERT(PT)	60.3	37.4
CoMVT(PT)	-	39.5
VQA-T* (PT)	-	41.5
MERRLOT(PT)	**69.5**	**43.1**
VGT (Ours)	61.6	39.7

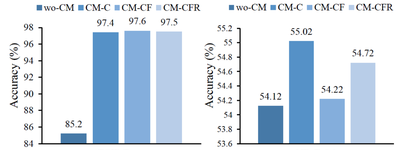


(a) NExT-QA (b) TGIF-QA (Trans)

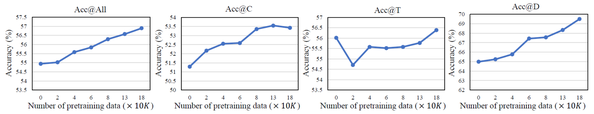
Classification model variant suffers from over-fitting.

- Ablation Study

Models	TGIF-QA		NExT-QA Val			
	Action	Trans	Acc@C	Acc@T	Acc@D	Acc@All
VGT	**95.0**	**97.6**	**52.28**	**55.09**	64.09	**55.02**
w/o DGT	89.6	95.4	50.10	52.85	64.48	53.22
w/o TTrans	94.0	97.6	50.86	53.04	64.86	53.74
w/o NTrans	94.5	97.4	50.79	54.22	63.32	53.84
w/o ETrans	94.8	97.4	51.25	54.34	64.48	54.30
w/o F_l	93.5	97.0	50.44	53.97	63.32	53.58
Comp→CLS	70.1	79.9	42.96	46.96	53.02	45.82



Cross-modal interaction at different levels.



(a) (b) (c) (d)

Conclusion:

- Contribution
  - We propose **video graph transformer** to advance VideoQA from coarse recognition and description to fine-gained visual reasoning in dynamic scenarios, and we achieve **SOTA** results on related benchmarks.
  - We propose **dynamic graph transformer** to encode visual graph dynamics for relation reasoning in space-time. Importantly, we demonstrate that **contrastive learning** significantly outperforms classification for multi-choice cross-modal video reasoning.
  - We are the 1st to shown that **pretraining visual graph transformer** can benefit video-language understanding towards a more data-efficient and fine-grained direction.