Video Graph Transformer for Video Question Answering - Supplementary Materials

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

A Data Statistics

The statistical details of the experimented datasets are presented in Table 1. For better comparison with previous works, we focus on the multi-choice QA task in NExT-QA [5] though it has also defined open-ended QA. For TGIF-QA [2], we also conduct experiments on a latest version [3] which generates more challenging negative answers for each question in the multi-choice tasks. In particular, we further fix the 'redundant answer' issue as we find that there are about 10% of questions have redundant candidate answers and some of the candidate answers are even identical to the correct one. The rectified annotations will be released along with the code.

Table 1. Data statistics. OE: Open-Ended QA. MC: Multi-Choice QA, VLen (s): Average video length in seconds.

Datasets	Main Challenges	#Videos/#QAs	Train	Val	Test	VLen (s)	QA
NExT-QA [5]	Causal & Temporal Interaction	5.4K/48K	3.8K/34K	0.6K/5K	1K/9K	44	MC
TGIF-QA [2]	Repetition Action	22.8K/22.7K	20.5K/20.5K	-	2.3K/2.3K	3	MC
	State Transition	29.5K/58.9K	26.4K/52.7K	-	3.1K/6.2K	3	MC
	Frame QA	39.5K/53.1K	32.3K/39.4K	-	7.1K/13.7K	3	OE
MSRVTT-QA [6]	Descriptive QA	10K/ 244K	6.5K/159K	0.5K/12K	3K/73K	15	OE

B Implementation Details

For training with QA annotations, we firstly train the whole model (except for the object detection model) end-to-end, and then freeze BERT to fine-tune the other parts of the best model obtained at the 1st stage. The best results in the two stages are determined as final results. Note that our hyper-parameters are mostly searched on the NExT-QA validation set and kept unchanged for other datasets. The maximum epoch varies from 10 to 30 among different datasets. For pretraining with data crawled from the Web, we randomly select 0.18M video-text data (less than 10%) from WebVid2.5M [4] [1]. The videos are then extracted at 5 frames per second and are processed in the same way as for QA. We then

[4] https://m-bain.github.io/webvid-dataset/

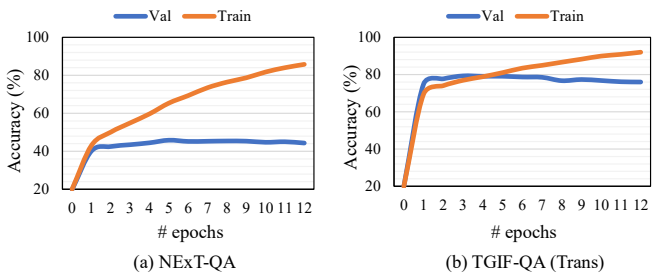

Fig. 1. Accuracy with regard to different training epochs.

optimize the model with an initial learning rate of 5×10^{-5} and batch size 64. The number of negative descriptions of a video for cross-modal matching is set to 63, and they are randomly selected from the descriptions of other videos in the whole training set. Besides, a text token is corrupted at a probability of 15% in masked language modelling. Following [7], a corrupted token will be replaced with 1) the '[MASK]' token by a chance of 80%, 2) a random token by a chance of 10%, and 3) the same token by a chance of 10%. We train the model by maximal 2 epochs which gives to the best generalization results, and it takes about 2 hours.

C Additional Model Analysis

C.1 Similarity Comparison *vs.* Classification

To study the reason for the poor performance of the classification model variant described in Sec. 4.3 of the main text, we visualize the training and validation accuracy with regard to different training epochs in Fig. 1. The results indicate that the classification model variant suffers from serious over-fitting issues, especially on NExT-QA [5] whose QA contents are relative complex but with less training data. To study whether the problem comes from the classification formulation or the cross-modal transformer, we further substitute the cross-modal transformer (CM-Trans) with our cross-modal interaction (CM) module introduced in Sec. 3.4 of the main text. We find that such a substitution can slightly alleviate the problem. For example, on NExT-QA val set, the accuracy increases from 45.82% to 46.98%. Nevertheless, the performance is still much worse than a comparison-based model implementation (*i.e.* 55.02%). This experiment reveals two facts: 1) Formulating QA problem as classification is the major cause for the weak performance. 2) The cross-modal transformer exacerbates the over-fitting problem, possibly because it involves additional parameters.

C.2 Study of Video Sampling

In Fig. 2, we study the effect of sampled video clips and region proposals on NExT-QA [5] test set. Regarding the number of sampled video clips, we find

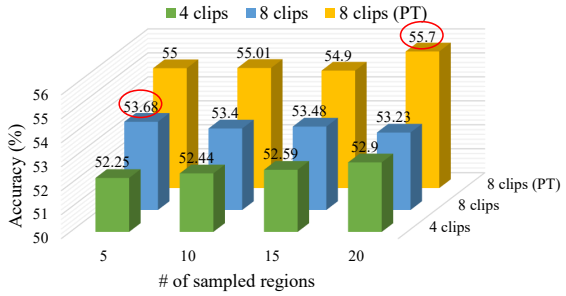

Fig. 2. Investigation of sampled video clips and region proposals per frame. Results are reported on NExT-QA test set.

Table 2. Comparison of memory and time based on NExT-QA [5]. (2m×8: 2 minutes per epoch and 8 epochs in total.)

Models	Acc@All	#Params (M)	GPU Memory		Time	
			Train	Infer	Train	Infer(FLOPs)
VQA-T [52]	45.30	156.5	5.6G	2.6G	2m×8	2448M
VGT (BERT)	55.02	133.7	16.2G	3.9G	7m×5	7121M
VGT (DistilBERT)	53.46	90.5	10.0G	3.5G	5m×7	3922M

that the setting of 8 clips steadily wins on 4 clips. This is understandable as the videos in NExT-QA are relatively long. As for the sampled regions, when learning the model from scratch, the setting of 5 regions gives relatively better result, e.g., 53.68%. Nonetheless, when pretraining are considered, the setting of 20 regions gives better result, e.g., 55.70%. Such difference could be due to that learning with more regions can yield over-fitting issues when the dataset is not large enough, since the constructed graph become much larger and more complex. Our speculation is also supported by the fact that the accuracy increases with the number of sampled regions when we only sample 4 video clips and thus less number of total graph nodes.

C.3 Model Efficiency

We compare VGT with VQA-T [7] in Tab. 2 for better understanding of the memory and time cost. Experiments are done on 1 Tesla V100 GPU with batch size 64. We use 1 example to report inference FLOPs. **Memory:** VGT has less training parameters (133.7M *vs.* 156.5M) and thus smaller model size than VQA-T (511M *vs.* 600M). The BERT encoder in VGT takes 82% of the parameters, the vision part is lightweight with only 24M parameters. VGT needs more GPU memory for training. Yet, the memory for inference are fairly small and close to that of VQA-T. We also implement a smaller version of VGT by replacing BERT with DistilBERT [4] as in VQA-T. With nearly 0.6× number of VQA-T's parameters (90.5/156.5M), we can still achieve strong performances (*i.e.* 53.46%). **Time:** Our FLOPs on 1 example is ∼2.9× that of VQA-T and ∼1.6× if we use DistilBERT. However, VGT converges much faster and needs much fewer epochs (total FLOPs) to get results superior to VQA-T when training with the

same data. For example, on NExT-QA, VGT's result at epoch 2 (50.16%) already significantly surpasses VQA-T's best result (45.30%) achieved at epoch 8. Also, VGT's result without pretraining can surpasses that of VQA-T pretrained with million-scale data. In this sense, VGT needs much fewer total FLOPs than VQA-T and other similar pretrained models for visual reasoning.

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1728–1738 (2021)
2. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2758–2766 (2017)
3. Peng, L., Yang, S., Bin, Y., Wang, G.: Progressive graph attention network for video question answering. In: ACM MM. pp. 2871–2879 (2021)
4. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. Advances in neural information processing systems (NeurIPS) (2019)
5. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9777–9786 (2021)
6. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: ACM MM. pp. 1645–1653 (2017)
7. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Just ask: Learning to answer questions from millions of narrated videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1686–1697 (2021)