The Lame Can't Go Far: Visual Stream Limits The Video Question Answering

Zuyao Chen[1,2], Jinlin Wu[2,3], Zhen Lei[2,3], Zhaoxiang Zhang[2,3], and Changwen Chen[1]

[1]The Hong Kong Polytechnic University.
[2]Centre for Artificial Intelligence and Robotics, HKISI, CAS.
[3]NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China.
zuyao.chen@connect.polyu.hk, {jinlin.wu, zlei}@nlpr.ia.ac.cn, zhaoxiang.zhang@ia.ac.cn,
changwen.chen@polyu.edu.hk

Abstract

*Video Question Answering (Video-QA) is a hot topic, which captures the present knowledge from surrounding situations and performs answer reasoning accordingly. Previous methods tackle this problem as a visual-language representation learning task. Benefiting from large-scale pre-training of NLP, the language stream of Video-QA makes a breakthrough. On the contrary, 2D or 3D ConvNets dominate the visual stream. The weak visual representation becomes the bottleneck of visual reasoning. To address this problem, we propose sparse-BEiT, a simple yet effective video-and-language representation learning framework. First, it decouples video representation learning with a divide-and-conquer strategy. Sparse-BEiT applies a strong pre-trained BEiT as the visual encoder to extract accurate visual representation from each frame. Then we adopt lightweight transformer layers to integrate temporal information. Moreover, to avoid drowning in the redundant information of the video, we introduce a temporal sparse sampling strategy, which samples a few frames from the video to encourage sparse-BEiT to guide the temporal aggregation module focusing on the temporal saliency clues. In this way, our sparse-BEiT ranks **1-st place** in the public STAR track of the first Machine Visual Common Sense: Perception, Prediction, Planning challenge on ECCV 2022. Our code will be released on* https://github.com/JosephChenHub/sparse-beit.git.

1. Introduction

Video-QA aims to develop an interactive AI system to communicate with the dynamic visual world via natural language. It looks forward to reasoning out the spatial, temporal, and causal relationships that are crucial for next-generation AI systems. Recently, the large-scale pre-training technology in NLP has made Video-QA a breakthrough, but the weak visual stream limits further development.

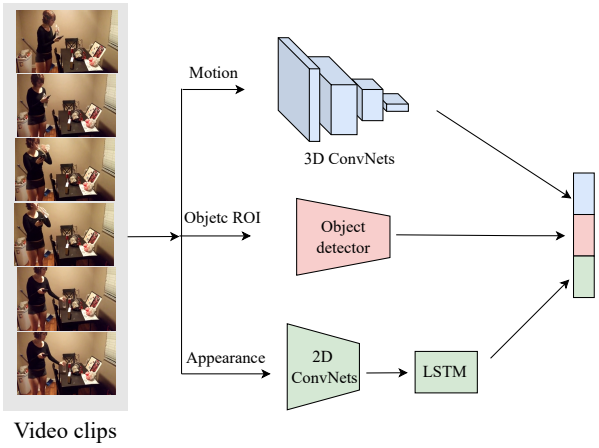


Figure 1. Existing visual representation methods in VideoQA. It commonly contains three branches, a 3D ConvNet for motion representation, a object detector for ROI and a 2D ConvNet with a LSTM for object appearance reprsentation.

As shown in Fig 1, existing Video-QA methods typically utilize a multi-granularity framework [9], consisting of a 3D ConvNet for motion representation, an object detector for ROI and a 2D ConvNet with LSTM for object appearance. However, in this approach, the 3D ConvNet models spatial-temporal information jointly, causing optimization difficulties. Object representation extracted by 2D ConvNets may not match the answer. Besides, the representation learned by different branches may not be in the same feature space, producing difficulties with branch integration. To solve this problem, we adopt a simple yet effective Video-QA frame-

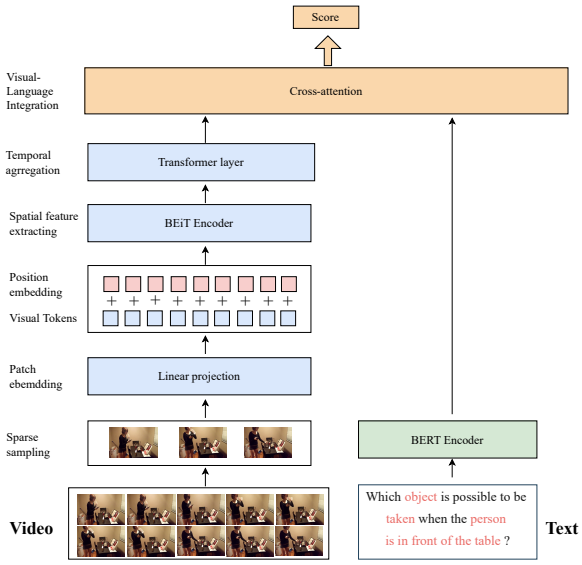


Figure 2. Framework of Sparse-BEiT contains a strong BEiT as the visual encoder, a two-layer transformer block as temporal aggregation module, a pre-trained BERT as the text encoder, and a visual-language integration module.

work named Sparse-BEiT in this paper. The framework of Sparse-BEiT is as shown in Fig. 2. First, we replace 2D ConvNets with a pre-trained strong BEiT [1] to obtain an accurate visual representation of objects. The BEiT is built with transformer layers without extra pooling or other spatial down-sampling operations. It preserves visual representations of objects in video clips and provides sufficient visual evidence for answer reasoning. Then, we introduce a two-layer transformer block as a temporal aggregation module, providing long-range modeling ability for the whole video. The spatial-temporal modeling problem is decoupled into a visual encoder (BEiT) and a temporal aggregation module in BEiT. Thus, the optimization becomes easier.

In addition, considering the redundancy of video frames, we introduce a sparse sampling strategy [4] that randomly samples sparse frames from the whole video clip to extract visual representations. Specifically, during the training stage, a few frames are randomly sampled from the video clip and fed into a vision encoder to obtain its visual representation, and text or question-answer pairs are fed into a language encoder to obtain the text feature. Then, these two multi-modal features are aggregated via a cross-attention module to generate the representation for video-language pairs.

2. Methodology

The overall framework is shown in Fig 2, which contains a vision encoder, a text encoder, and a visual-language integration module.

Visual encoder. In this paper, we adopt the naive ClipBERT [4] as our baseline model. Although ClipBERT [4] achieves a trade-off between the computation and performance, the weak visual encoder and the misalignment of different modality features limit the reasoning capability. To solve this, we replace the ResNet-50 [3] with a strong vision transformer, BEiT, which is pre-trained on the large-scale ImageNet-22k dataset. The large-scale pre-training makes BEiT have a stronger scalable ability in downstream vision tasks than other ConvNets. The absence of down-sampling operations in transformer layers ensures that BEiT is able to extract finer-grained features that provide sufficient visual evidence for answer reasoning.

Visual-language integration. We use BERT [2] acts as the text encoder, and a lightweight cross-attention module is adopted to aggregate the multi-modal information.

Formally, given a video-text pair as V (for video) and T (for text sequence), the video V is further denoted as a list of N clips $[c_1, c_2, \cdots, c_N]$. This standard paradigm can be formulated as:

$$p = \mathcal{F}(E_t([E_v(c_1), E_v(c_2), \cdots, E_v(c_N)]), E_l(T)) \quad (1)$$

where $\mathcal{F}$ denotes the visual-language integration module. $E_l(\cdot)$ and $E_v(\cdot)$ are text encoder and visual encoder. E_t denotes the temporal aggregation module. For STAR challenge, we formulate the question answering as a multiple choice problem, and cross-entropy loss is adopted as follows for training:

$$\mathcal{L} = -\frac{1}{|\mathcal{D}|}\sum_{n=1}^{|\mathcal{D}|}\log(p_n[y_n]) \quad (2)$$

where $|\mathcal{D}|$ denotes the number of dataset, p_n is the logit after softmax, and y_n is the index of ground-truth.

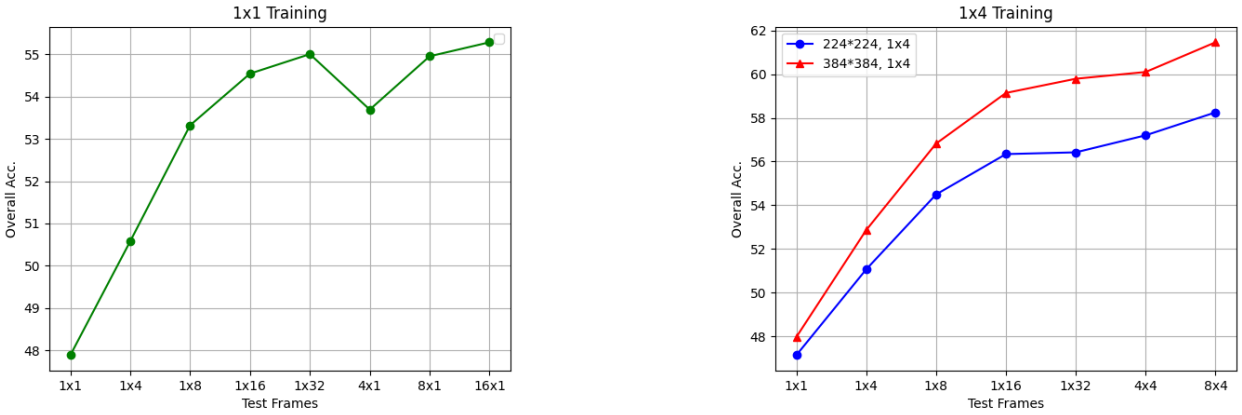
3. Experiments

Dataset. The STAR challenge is built on the Charades dataset [7], which contains video clips, and the average duration of the video is about 10 seconds. The annotation of the STAR challenge has been split into three parts, *i.e.*, $45,731$ clip-question pairs for the training set, $7,098$ pairs for the validation set, and $7,377$ pairs for the test set. During implementation, we use PyAv to decode the compressed video.

Implementation details. The framework is implemented with PyTorch, and the BEiT-base model is adopted as the vision encoder, a two-layer transformer block is used as the temporal aggregation module, the BERT base model is used as the text encoder, and three attention layers are used as the cross-attention module for multi-modal information aggregation. Due to there being four choices for each question in STAR, we formulate the video question answering as a

Method	Input resolution	Training Frames	Test Frames	Interaction	Sequence	Prediction	Feasibility	Overall Acc.
ClipBERT	224*224	8x2	16x16	43.62	48.35	41.83	45.51	45.99
Sparse-BEiT	224*224	1x1	16x1	52.87	57.92	52.56	51.22	55.28
Sparse-BEiT	224*224	1x4	8x4	55.92	61.41	53.04	53.27	58.25
Sparse-BEiT*	384*384	1x4	8x4	59.13	64.39	58.01	55.92	61.46

Table 1. Performance on validation set of STAR, where "$M \times N$" refer to M clips and each clip has N frames, ClipBERT use ResNet-50 [3] as the backbone, and Sparse-BEiT, Sparse-BEiT* use BEiT [1] as the backbone.



(a) (b)

Figure 3. #frames at inference on the validation set. (a) Training with 1×1; (b) Training with 1×4.

multiple-choice problem. We use AdamW [5] to optimize the end-to-end model, with maximum learning rate $1e-5$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and use a warm-up learning rate scheduling. The model is trained for 10 epochs with mixed precision on 4-A100 GPUs with batch size 64.

Comparison with SOTA. We report the quantitative results on the STAR validation set in Tab. 1. By replacing the vision encoder ResNet-50 [3] with BEiT [1], we can see a significant improvement with less training and testing frames, which **indicates that the visual features are essential for vision-language understanding**. Increasing the training frames from 1×1 to 1×4 brings us a 3 points enhancement; this is due to only one sampled frame is not enough to represent the whole frame state, and more frames information is needed, especially for the action that requires long-range dependencies. Moreover, by improving the input resolution, the overall accuracy has been improved about 3 points. However, higher input resolution means a higher computation and memory burden, so a trade-off exists between the accuracy and computation or memory.

We use the ensemble strategy in the inference stage as ClipBERT [4] does. From Fig. 3, we can see that even training with 1×1, the performance can be largely enhanced by aggregating more frames predictions, *e.g.*, 16×1 vs. 1×1. Another observation is that increasing input resolution benefits the inference that is consistent with with Tab. 1.

4. Discussion

Dense-to-sparse. To reduce the redundancy of modeling spatial-temporal sequences, two schemes can be considered. First, sparse sampling frames like ClipBERT [4] can be used to train the model and the ensemble strategy is adopted in the inference stage.

This method requires fewer modifications from image-based vision-language understanding. However, the undersampling of dense frames may cause a misalignment between video clips and the text, *e.g.*, same frames are sampled for different actions. Another dense-to-sparse strategy is proposing import tokens like [6].

Sparse-to-dense. For long-time video understanding, the frames will increase dramatically, and thus it requires a more efficient modeling strategy. One possible direction is to use a few key frames to estimate the full clip information. A similar application can be referred to accelerating compressed video object segmentation [8], which leverages the Codec information to propagate the segmentation mask for accelerating inference.

5. Conclusion

In this work, we find that the weak visual stream limits the development of current Video-QA. We enhance the visual stream by using a powerful vision transformers BEiT [1], introducing a sparse frame sampling strategy and in-

creasing the input resolution. In this way, our model achieves state-of-the-art performance in the public STAR dataset and ranks 1-st place in the public STAR track of the first Machine Visual Common Sense: Perception, Prediction, Planning challenge.

Acknowledgement

This work is supported by InnoHK program.

References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[4] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021.

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019.

[6] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

[7] Gunnar A Sigurdsson, Olga Russakovsky, and Abhinav Gupta. What actions are needed for understanding human actions in videos? In *Proceedings of the IEEE international conference on computer vision*, pages 2137–2146, 2017.

[8] Kai Xu and Angela Yao. Accelerating video object segmentation with compressed video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1342–1351, 2022.

[9] Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*, 2022.