Vision And Language Final Report

Yu-Wei Su
yfs5313@psu.edu

1. Task

The task of the paper is to tackle Video Question Answering(VideoQA). The goal of VideoQA is to answer questions about the content of a given video. It involves understanding the visual scenes in the video as well as the linguistic semantics of the question to produce an accurate answer. VideoQA is more complex than ImageQA (image-based question answering) because it requires the model to understand the temporal dynamics and relationships within the video content. [16]

The paper proposes a new long-form Video Question Answering (VideoQA) model called multi-modal iterative special-temporal Transformer (MIST). Previous models tackling VideoQA often focus on only short or video clips (1-2 seconds). However, some challenges arise when it comes to longer videos. First, Long videos contain more events thus requiring the models to have more complex temporal reasoning to answer the question. Traditional methods either rely on dense, computationally expensive video sampling or sparse sampling, which may overlook crucial visual information. Second, the question may contain objects, events, and relations that happen at other times compared to the simultaneous events in short videos.

2. Related Work

2.1. Video Question Answering SOTA (in AGQAv2)

As listed in https://paperswithcode.com/dataset/agqa. Two current competing SOTA are in the VideoQA task, especially in the AGQA dataset. The first is MIST, and the second is [3], which can be seen from the table in fig1. Although [3] are winning in half of the sub-categories, including overall AGQA performance. MIST is still better in temporal-related questions such as duration and sequencing tasks.

3. Approach

MIST decomposes the traditional dense spatial-temporal self-attention into cascaded segment and region selection modules that adaptively select frames and image regions relevant to the question. 2) Iterative Selection and Attention

Question Types	Most Likely	PSAC	HME	HCRN [23]	AIO [40]	Temp[ATP] [5]	MIST - AIO
Object-relation	9.39	37.84	37.42	40.33	48.34	50.15	51.43
Relation-action	50.00	49.95	49.90	49.86	48.99	49.76	54.67
Object-action	50.00	50.00	49.97	49.85	49.66	46.25	55.37
Superlative	21.01	33.20	33.21	33.55	37.53	39.78	41.34
Sequencing	49.78	49.78	49.77	49.70	49.61	48.25	53.14
Exists	50.00	49.94	49.96	50.01	50.81	51.79	53.49
Duration comparison	24.27	45.21	47.03	43.84	45.36	49.59	47.48
Activity recognition	5.52	4.14	5.43	5.52	18.97	18.96	**20.18**
All	10.99	40.18	39.89	42.11	48.59	49.79	50.96

Table 1. QA accuracies of state-of-the-art (SOTA) methods on AGQA v2 test set.

Figure 1. AGQA v2 performance comparison captured from [3]

over Multiple Layers: MIST iteratively conducts selection and attention over multiple layers to support reasoning over multiple events, which can be shown in fig2. Experimental results on four VideoQA datasets demonstrate that MIST achieves state-of-the-art performance while being computationally efficient and interpretable. A more detailed ex-

[hbtp]

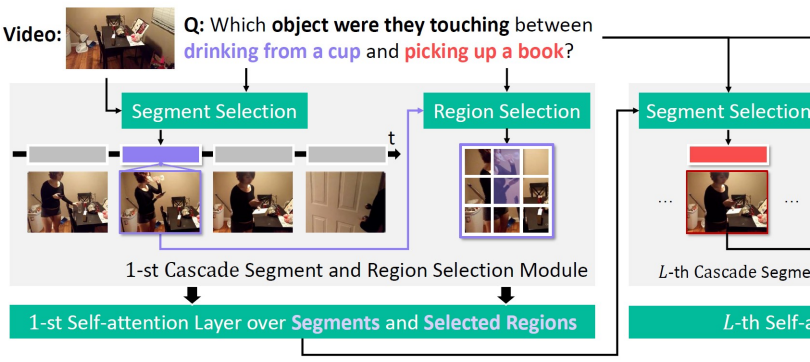

Figure 2. Model diagram as from Original paper

planation of its method can be seen in figure 3. The key part of MIST is that it tries to perform attention to segments of both temporal and regional parts. First, it will dissect a video into k segments, pick one patch by pooling, and then perform attention to each layer's selected patch. After that, the selector will select TOP_k result from k parts to get the temporal features. The model will further perform region selection on the TOP k results. The region section can

also be done using a similar approach by dissecting regions into top j results. With spatial, temporal, and word features obtained, multi-head attention is then performed to capture cross-related relations.

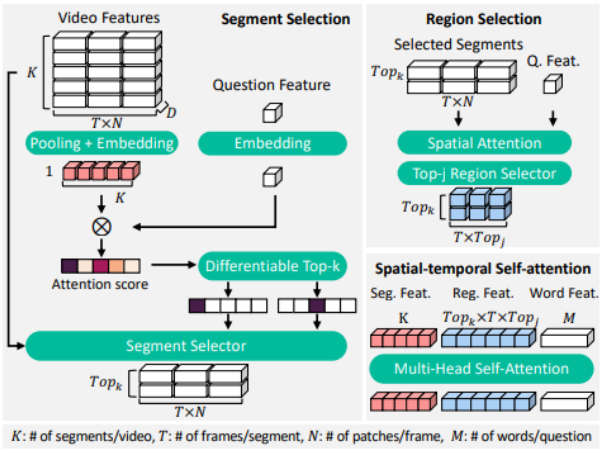

Figure 4. **Key components of Iterative Spatial-Temporal Attention Layer.** Since region selection follows the same architecture as segment selection, we only show its inputs and outputs.

Figure 3. Figure 4 from [7]

4. Dataset

The datasets [7] has trained on are in figure 4, which is Table 1 captured from [7]. This report focuses on replicating AGQA v2 [8], specifically. As [7] has stated, AGQA is an open-ended VideoQA benchmark for compositional spatio-temporal reasoning. The V2 version has a more balanced distribution. It provides 2.27M QA pairs over 9.7K videos with an average length of 30 seconds.

5. Results

There is no pretrained model provided by the original paper but only with the results from the paper. Thus, I only reported a comparison between my replication result and the paper's result. A comparison between the paper's result and my replication result is shown below. As we can see,

Table 1: Comparison of the Multi-Event VideoQA benchmarks.

Datasets	# Avg. Events/Video	Video Length	#Event Classes	#Videos	# QAs
STAR[42]	2.7	12s	157	22K	60K
AGQA v2[14]	6.8	30s	157	9.7K	2.27M
EgoTaskQA[16]	5	25s	793	2K	40K
NExT-QA[43]	8.8	44s	50	5.4K	52K

Figure 4. Enter Caption

Model	original	epoch 5	epoch 10	epoch 14	epoch 16
Accuracy	54.39	51.20	52.88	53.85	53.62

Table 1. replication vs original result.

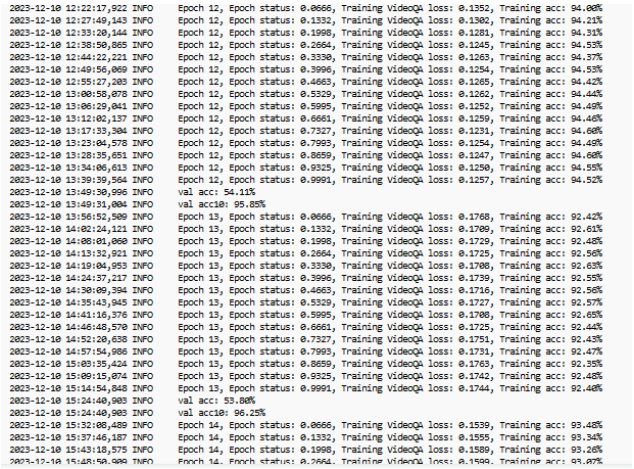

Figure 5. portion of training log-1.

The model has acquired a decent already in the first few epochs.

5.1. Issues

1. The replication is only performed to Epoch 16 due to the limitation of the computing power.

2. The first few training epochs are trained on the same learning rates. In the first few epochs, the connection of collab has been bad, so the training kept stopping. The fact that MIST uses a learning rate schedular was later discovered in the middle of the training process. So only epochs trained from 11 to 16 are modified and trained on the same scheduled learning rate same to the original result.

3. The replication results are reported with the val dataset due to the reason that the author does not provide the testing set data.

5.2. Screenshots of training log

Some portions of the training log of MIST are provided below (5, 6).

6. Possible Improvements and Results

6.1. Possible Improvements

There are two possible improvements, but I only tried to improve one. First is the fixed segments, and the other is the optimizer-related potential improvement.

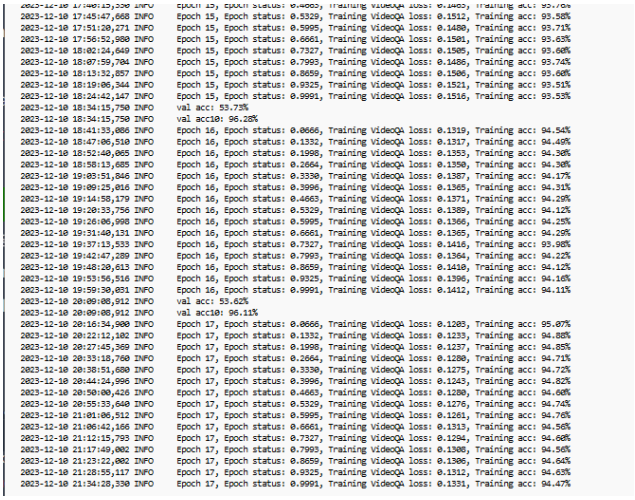

Figure 6. Enter Caption

6.1.1 fixed segments

it is unknown whether the model will perform poorly if the events are cut into different pieces. It may miss some vital event features due to the uniform segments without understanding segments of interest. Also, fixing K will lead the model to have a longer length per segment as the video lengths become longer. It could lead to a worse performance as irrelevant information is increased per segment.

Potential improvements could be tweaking the models to adaptively choose the hyperparameter K or cutting video segments into non-uniform segments to retain event information. Or even different K segments could also be used to capture long-term events and short-term events. Some challenges like computational expansiveness and how to feed it with different sizes would also arise. Due to the time limitation, I did not try to solve fixed segment issues.

6.1.2 Sharpness-Aware Minimization

Many works suggest that sharpness can be positively correlated with test errors of DNN ([15], [17], [14], [4], [12]). However, the following works showed that this correlation with generalization is not strong enough for Standard Sharpness. Some work [5] also shows that SAM could be helpful in vision and language tasks. Even though there are pieces of empirical evidence that don't agree with it [9], [10], [1]. I still would like to test whether it would help the generalizations.

The sharpness can be formulated as follows([2]):

$$s(\omega, S) := \max_{\|\delta\|_2 \leq \rho} \frac{1}{|S|} \sum_{i:(x_i,y_i)\in S} \ell_i(\omega+\delta) - \ell_i(\omega)$$

where $S = \{x_i, y_i\}_{i=1}^n$ be the training set, and $\ell_i(\omega)$ be the loss function corresponding to weights $\omega \in R^d$. The

Model	original	SAM epoch 2	SAM epoch 6
Accuracy	54.39	51.76	51.17
Model	MIST epoch 5	MIST epoch 6	MIST epoch 14
Accuracy	51.20	51.28	53.85

Table 2. replication vs original result.

key is that if one solution is sharper than the other solution, the maximum of its neighborhoods will have larger values since it goes up more drastically. It is seen that a solution with higher sharpness has lower generalization compared to a flatness solution empirically [6] [13]. Thus a Sharpness-Aware Minimization is proposed to minimize the object function with the care of sharpness.

With $\ell(\omega)$ be the original Empirical Risk Minimization(ERM) loss:

$$\ell_(\omega) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(\omega)$$

The SAM objective for perturbation would be:

$$\ell^{SAM}(\omega) = \max_{\|\delta\|_p \leq \rho} \frac{1}{|n|} \sum_{i=1} \ell_i(\omega+\delta)$$

An adaptive sharpness measure, which is what I tested, is also proposed to tackle rescaling issues [11]. The SAM package I used can be found at https://github.com/davda54/sam.

6.2. Modification Results

The SAM-modified version results are compared with the Original and non-modified versions. A comparison between the paper's, non-modified replication results, and SAM-MIST is shown below in table 2. The SAM-MIST actually has a better validation accuracy within a few epochs. However, since SAM requires two back propagation computing, the training time of the SAM version is twice for an epoch. Thus, only six epochs are trained on SAM, and whether it could perform better in the following epochs is unknown.

6.3. Issues

1. The replication is only performed to Epoch 6 due to the limitation of the computing power.

2. Similar to the MIST training. The first few training epochs are trained on the same learning rates. In the first few epochs, the connection of collab has been bad, so the training kept stopping. The fact that MIST uses a learning rate schedular was later discovered during the training process. So Epoch 6 may actually be trained on the learning rate from Epoch 4 of the original settings.

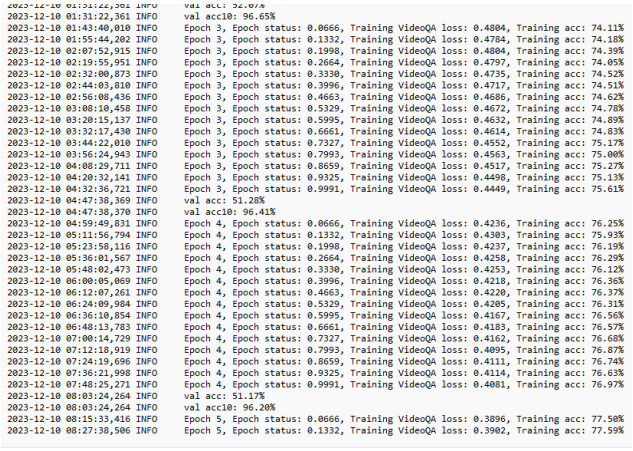

Figure 7. log of SAM-MIST training

3. The replication results are reported with the val dataset because the author does not provide the testing set data.

6.4. Screenshots of training log

Some portions of the training log of SAM-MIST are provided below (7,).

7. Code Repository

The link to my training code can be found in. The original MIST github can be found in https://github.com/thejackys/SAM-MIST. The SAM optimizer can be found in https://github.com/davda54/sam. I only included the modified version for SAM as the replication version is basically follows the MIST's own Github repo https://github.com/showlab/mist.

References

[1] Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. *arXiv preprint arXiv:2302.07011*, 2023. 3

[2] Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization, 2022. 3

[3] Ziyi Bai, Ruiping Wang, and Xilin CHEN. Glance and focus: Memory prompting for multi-event video question answering. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1

[4] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021. 3

[5] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pretraining or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021. 3

[6] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization, Apr 2021. 3

[7] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering, 2022. 2

[8] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa 2.0: An updated benchmark for compositional spatio-temporal reasoning, 2022. 2

[9] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019. 3

[10] Simran Kaur, Jeremy Cohen, and Zachary Chase Lipton. On the maximum hessian eigenvalue and generalization. In *Proceedings on*, pages 51–65. PMLR, 2023. 3

[11] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pages 5905–5914. PMLR, 2021. 3

[12] Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. *Advances in Neural Information Processing Systems*, 35:34689–34708, 2022. 3

[13] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning, 2017. 3

[14] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 3

[15] Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with sgd. *arXiv preprint arXiv:1802.08770*, 2018. 3

[16] Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges, 2022. 1

[17] Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, et al. Towards theoretically understanding why sgd generalizes better than adam in deep learning. *Advances in Neural Information Processing Systems*, 33:21285–21296, 2020. 3