Supplementary: Watching the News: Towards VideoQA Models that can Read

Soumya Jahagirdar[†] Minesh Mathew[†] Dimosthenis Karatzas[‡] C. V. Jawahar[†]
{soumya.jahagirdar, minesh.mathew}@research.iiit.ac.in dimos@cvc.uab.es jawahar@iiit.ac.in
[†] CVIT, IIIT Hyderabad, India [‡] Computer Vision Center, UAB, Spain

1. Screenshot of Annotation tool.

We collect the NewsVideoQA dataset using the annotation tool. The screen shots for the annotation tool are shown in Fig. 1 and Fig. 2.

2. Detailed experimental setup OCR-aware SINGULARITY

OCR-aware SINGULARITY. The model is implemented in PyTorch [6]. Similar to how SINGULARITY was originally implemented, we initialize vision encoder using BEiT_BASE model which was previously trained on ImageNet-21K [2]. Using the NewsVideoQA dataset, we further pretrain the previously trained SINGULARITY (checkpoint: singularity_temporal_17m.pth). Original SINGULARITY uses a video/image-caption combination for pretraining. Instead, we combine the video with OCR tokens of the frames corresponding to the timestamp of the questions defined. Similar to original setting, in pretraining, a single frame is sampled from the whole video. We pretrain the model for 10 epochs.

For the task of finetuning, we concatenation the question tokens with the OCR tokens with the OCR tokens of the frame on which the question is defined. An additional multimodal decoder is initialized from pretrained multimodal encoder (pretrained on NewsVideoQA dataset). It uses the multimodal encoder outputs as cross-attention inputs and decodes the answer with start token as [CLS]. In the first half period, the learning rate is $1e-4$ with a warm up factor, followed by cosine decay to $1e-6$. We finetune the model for 20 epochs with a batch size of four. When training the model, a single frame is used, and when testing the model, 12 frames are used. At the time of inference, similar to [5], we use early fusion strategy, which takes all frames as model inputs (concatenation of all the frames considered at the test time) for directly making a more informative video-level prediction.

3. Quantitative Results for M4C with two frames setup

We report the findings for M4C [4] (previously trained M4C on TextVQA [7] and ST-VQA [1]) and BERT-QA [3] with a single frame while training, and tested on two frames per question in Table. 2, and Table. 1. Because the frames are captured at 2 frames per second when preprocessing videos to obtain the response for OCR tokens, we may acquire at most two frames per second for a question defined at a specific timestamp. We utilized the first frame in the single frame experiments, and we utilize both frames corresponding to the timestamp of the question for the two frames experiments.

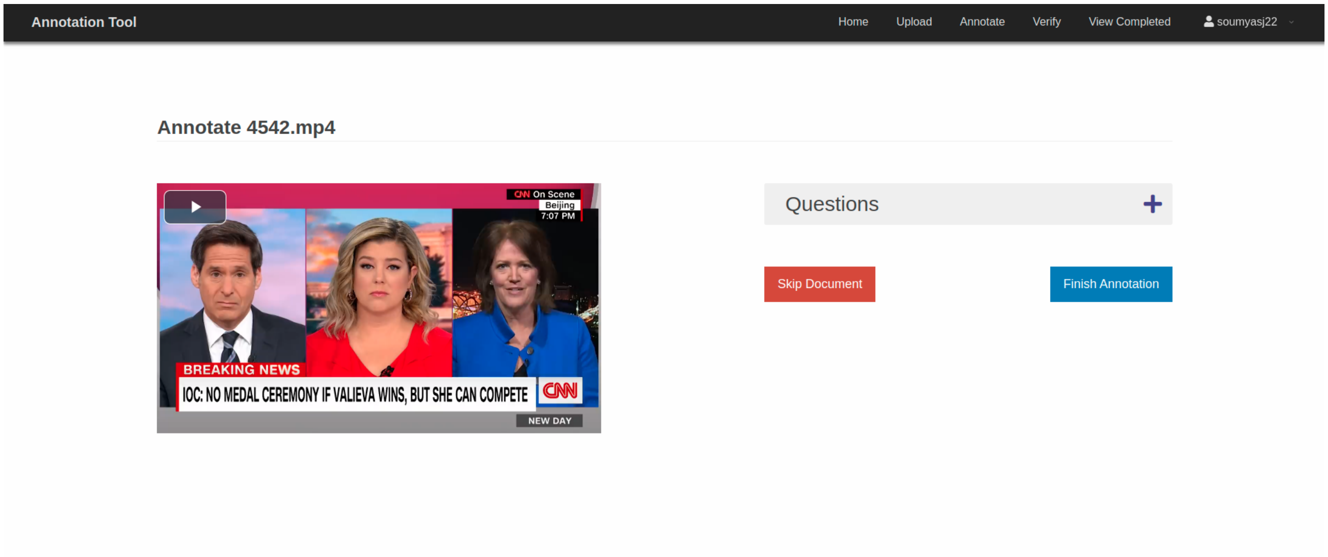
Calculation of accuracy and ANLS for two frames as test input. For BERT [3], and [4], we check if the ground truth answer is present in either of the two answers predicted by the two frame-question pair to obtain the final answer for each question. (At test time: input is two frames each paired with question.)

INPUT: [frame1 + Question], OUTPUT: [Answer1]
INPUT: [frame2 + Question], OUTPUT: [Answer2]

OUTPUT: if ground truth answer is present in either of the two answer predictions, [Answer1, Answer2], then it is a correct prediction. To calculate ANLS, we separately calculate ANLS for each answer (Answer1, Answer2) and average the score for both.

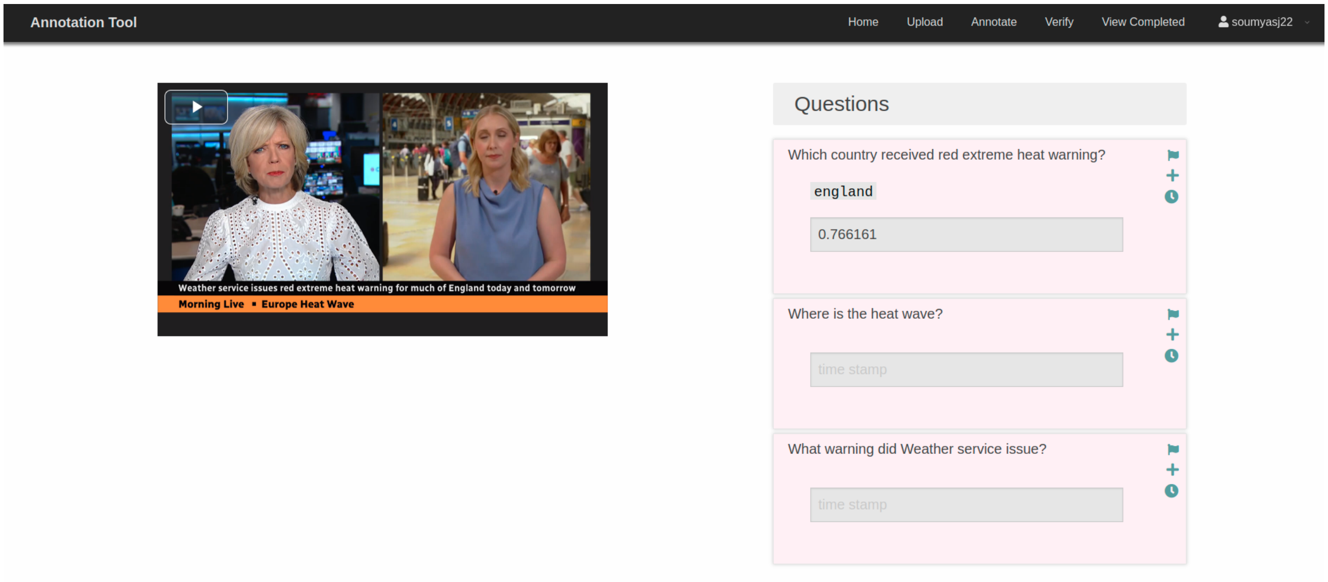
BERT. In Table. 1, we show the comparisons for two cases: (i) SF: Single Frame, and (ii) TF: Two frames. In single frame, we use OCR tokens of correct frame corresponding to the timestamp of the question. In two frames, we concatenate OCR tokens from two frames corresponding to the timestamp of the question.

M4C. In Table. 2, we show the results of M4C trained on two different settings. (i) Image features are included as input to M4C, and (ii) Image features are set to zero. We also show results for M4C trained from scratch and pretrained M4C (on TextVQA+ST-VQA). The difference in the performance indicates the less dependency of models output



Annotate 4542.mp4

Figure 1: Annotation tool used for annotating QA pairs. General setup for annotation tool.


Annotation Tool
Home
Upload
Annotate
Verify
View Completed
soumyasj22
Weather service issues red extreme heat warning for much of England today and tomorrow
Morning Live ▪ Europe Heat Wave
Questions

Which country received red extreme heat warning?

england

0.766161

Where is the heat wave?

time stamp

What warning did Weather service issue?

time stamp

Figure 2: Annotation tool with questions, answers, and respective timestamp.

on static visual features. Results for M4C with two frames are in the supplementary.

SINGULARITY and OCR-aware SINGULARITY. (i) We show the performance of SINGULARITY finetuned on NewsVideoQA with [video-question-answer] as input. (ii) We show the performance of SINGULARITY finetuned on NewsVideoQA with [video-question-answer-OCR tokens] as input. (iii) We compare the above mentioned cases with OCR-aware SINGULARITY which is pretrained on NewsVideoQA dataset and finetuned on NewsVideoQA with [video-question-answer-OCR tokens] as input. More details and explanations are in given in supplementary.

Using Tables. 1, 2, and 3 we summarise the performance of different methods on different tasks. Scene text-aware models like M4C are designed for VQA task that require textual understanding to answer the questions on a single image, hence performs well when the correct frame is used as input. Video question answering methods like SINGULARITY consider only video features to obtain answers to the questions. These models do not consider scene-text/embedded text in the videos, therefore perform poor on text-based VideoQA motivating us to add OCR tokens as

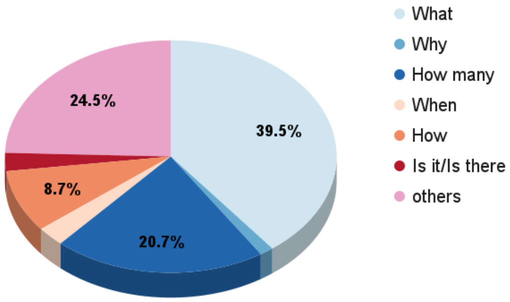


Figure 3: Statistics about NewsVideoQA dataset based on type of questions. Note that there is diverse range of types of questions in the dataset. The question type *"What"* has a maximum count with questions such as, *"What could be the reason ...?", "What is the value..?", "What is one of the..?"* and so on.

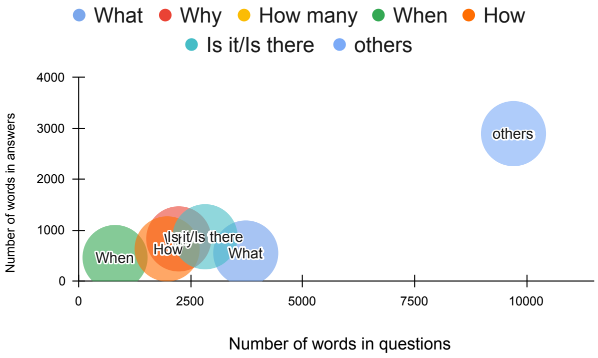


Figure 4: A bubble chart with number of words in question in horizontal axis and number of words in answers in vertical axis based on question type.

an extra input that significantly improves the performance of SINGULARITY.

It should be noted that, BERT-QA and M4C are single frame models where the correct frame on which the question is asked, is assumed to be provided. But SINGULARITY is a video QA model that finds the correct answer by looking at more than one frame. It can be seen that, M4C achieves the highest performance as it was previously trained on Scene text aware VQA datasets like TextVQA [7] and ST-VQA [1]. To have a fair comparison i.e training with a single frame (like singularity) and testing on multiple frame setting, we conduct experiments by modifying BERT-QA and M4C. These models are given 12 frames while testing, and the final answer is obtain by majority voting. Results for this experiment are presented in the main paper.

Table 1: **Results on BERT-QA.** BERT-QA-SF represents BERT-QA trained on a context being OCR tokens from one frame per QA pair. BERT-QA-TF represents BERT-QA trained on OCR tokens from two frames per QA pair.

Model	#Frame	Finetuning	Acc. (%)	ANLS
BERT-QA-SF	1	✗	22.96	29.03
BERT-QA-SF	1	✓	28.70	34.21
BERT-QA-TF	2	✗	24.42	30.28
BERT-QA-TF	2	✓	**31.94**	**36.94**

Table 2: **Results on M4C.** M4C_scratch is trained only on NewsVideoQA. In case of M4C_finetuned, an M4C model that is already trained on Text VQA and ST-VQA is finetuned on the NewsVideoQA. Since all questions are grounded on textual information, visual objects such as people or things seen in the video provide little information for answering the questions.

Model	visual objects	Acc. (%)	ANLS
M4C_scratch	✗	31.83	36.21
M4C_scratch	✓	28.70	33.53
M4C_finetuned	✗	**35.073**	**39.62**
M4C_finetuned	✓	28.49	32.17

Table 3: **Results on SINGULARITY and OCR-aware Singularity.** OCR-aware singularity performs the best.

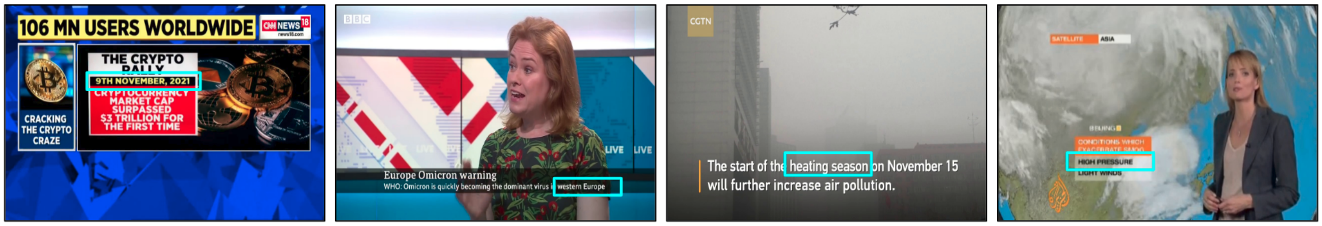
Model	OCR	Pretraining	Acc. (%)	ANLS
SINGULARITY	✗	✗	4.82	5.78
SINGULARITY	✓	✗	31.38	35.27
SINGULARITY	✓	✓	**33.57**	**37.52**

4. Examples from the dataset

In Fig. 4, we show the distribution of questions based on their types derived by checking presence of question words such as "What", "Why", "How many" and so on. In Fig. 3, shows the distribution of the statistics about NewsVideoQA dataset based on the type of questions. These figures indicate the variability in the dataset with respect to different types of questions.

In Fig. 2 and Fig. 6, we show some examples from the dataset. In Fig. 7 we show an example from the dataset that requires textual information from multiple frames to answer a question and that can be answered using the visual cues. The textual cues helps the inference made using visual cues only thereby resulting in accurate answer predictions in VideoQA.

Along with the supplementary pdf, we have added a folder named "sample_video". It contains a sample video from the dataset named: 383x3.mp4, annotation json file name: 383x3_sample.json which contains question, answer, timestamp and other annotations like video title, video link, and ocr_information folder: which contains information of


106 MN USERS WORLDWIDE
THE CRYPTO RALLY
9TH NOVEMBER, 2021
CRYPTOCURRENCY MARKET CAP SURPASSED $3 TRILLION FOR THE FIRST TIME
CRACKING THE CRYPTO CRAZE
Europe Omicron warning
WHO: Omicron is quickly becoming the dominant virus in western Europe
The start of the heating season on November 15
will further increase air pollution.
HIGH PRESSURE

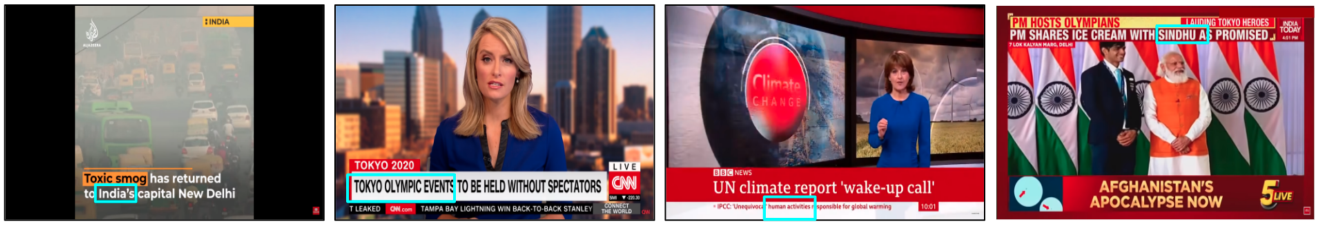
Question:
When did crypto market surpass 3 trillion dollar for the first time?
Answer:
9th november 2021

Question:
Where is the omicron variant becoming dominant?
Answer:
western europe

Question:
Start of what will increase air pollution?
Answer:
heating season

Question:
What is the first condition that causes smog?
Answer:
High pressure

Figure 5: Examples of NewsVideoQA dataset showcasing the importance of text in the videos to answer questions. Examples from multiple topics are shown to indicate the variability in the type of questions.


INDIA
Toxic smog has returned
to India's capital New Delhi
TOKYO 2020
TOKYO OLYMPIC EVENTS TO BE HELD WITHOUT SPECTATORS
CNN
LIVE
UN climate report 'wake-up call'
Climate CHANGE
PM HOSTS OLYMPIANS
PM SHARES ICE CREAM WITH SINDHU AS PROMISED
AFGHANISTAN'S APOCALYPSE NOW

OCR tokens:
... toxic smog has returned to india's capital ...
Question:
Which country is facing toxic smog?
Answer:
india

OCR tokens:
... tokyo 2020 tokyo olympic events to be held without spectators ...
Question:
Which event is held without spectators?
Answer:
tokyo Olympic events

OCR tokens:
... bbc news un climate report 'wake-up call' ipcc ...
Question:
What are responsible for global warming?
Answer:
human activities

OCR tokens:
... pm hosts olympians pm shared ice cream with sindhu as promised ...
Question:
Who shared Ice Cream with the PM?
Answer:
sindhu

Figure 6: More examples from NewsVideoQA dataset.

OCR tokens from the frames that were sampled at 2FPS.

References

[1] Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez i Bigorda, Marçal Rusiñol, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pages 4290–4300. IEEE, 2019.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[4] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *CVPR*, pages 9989–9999. Computer Vision Foundation / IEEE, 2020.

[5] Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning, 2022.

[6] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019.

[7] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

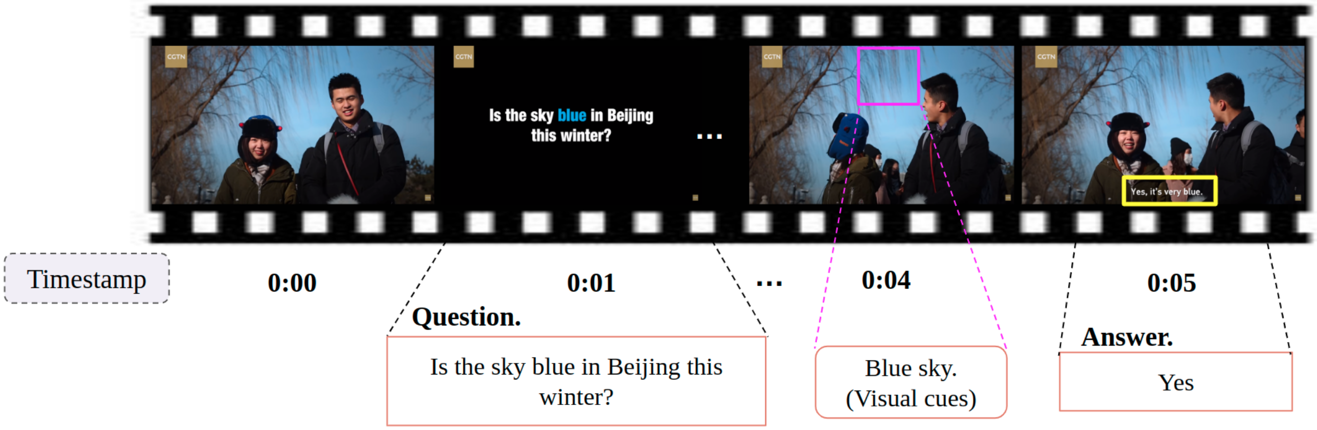


Figure 7: Examples of NewsVideoQA dataset showcasing the examples where a question is framed at timestamp of 0:01 in a video. The answer to this question is in timestamp of 0:05. Also, this question can be answered by visual cues in the video, but the text at 0:05 timestamp confirms the inference made by text.