Supplementary: Understanding Video Scenes through Text: Insights from Text-based Video Question Answering

Soumya Jahagirdar[1] Minesh Mathew[2] Dimosthenis Karatzas[3] C. V. Jawahar[1]
soumya.jahagirdar@research.iiit.ac.in minesh@wadhwaniai.org dimos@cvc.uab.es jawahar@iiit.ac.in
[1] CVIT, IIIT Hyderabad, India [2] Wadhwani AI [3] Computer Vision Center, UAB, Spain

As discussed in the main paper, we add extra annotations for each question-answer pair a type for M4-ViteVQA [3] dataset. These are as follows:

1. Extractive-based: These are the type of answers which can be directly extracted from the context which in this case is OCR tokens. (Meaning that the answer is a substring of the concatenated string of OCR tokens)

2. Reasoning-based: For this type, the answer can be obtained by reasoning over the content in the image (both visual and textual). The examples could be yes/no type of questions where answer is not always present in the text or visual content but it needs the model to reason over multiple modalities.

3. Knowledge-based: In this type, the answers usually require external knowledge such as knowing that a particular brand name is for a specific product.

1. Quantitative results

Tab. 1 presents the results of the BERT-QA [1] model on M4-ViteVQA [3] dataset. The evaluation consists of a comprehensive series of experiments. We divide the experiments into two types. In the first type, we consider all the questions from the validation set for all the original splits i.e. `Task1Split1`, `Task1Split2`, and `Task2`. For the second type, we narrow down our analysis to only include question-answer pairs where the answer is found within the context. In both Tab. 1 and 3, it can be seen that we only the results on BERT-QA model for the cases when for the cases where answer is present in the context (for the entire validation set, the results are already shown and compared to T5-ViteVQA in the main paper) because the code for T5-ViteVQA or the pretrained models are not open sourced.

Specifically, we extract the concatenated list of OCR tokens from frames sampled at 1fps and consider only those pairs where the answer appears in this context. It can be seen that with finetuning the performance of BERT-QA increases over all tasks and splits. It can also be observed

Question: How much is Bitcoin?

Answer: $44,375.88

Sets: {Single frame, Visual}

Figure 1. A specific example from M4-ViteVQA [3] dataset. For this example, the dataset provides two annotations, (1) whether the question can be answered by a single frame, and (2) whether a question needs visual information along with textual information to obtain answer.

that the ANLS metric is significantly better than accuracy. This is due to the fact that the OCR tokens do not have ground truth default reading order which is very essential for extractive question-answering models like BERT-QA. It is also because the OCR annotations provided in M4-ViteVQA [3] dataset are obtained from open-sourced OCR detection and recognition, which can be improved by using a commercial OCR such as GoogleOCR. In the second type of experiment, where we evaluate only the questions that have answers in the context, we observe improved scores. This outcome is somewhat expected, given that we are specifically selecting question-answer pairs where the answer is present in the context. Consequently, the model's performance is naturally better in this scenario.

In Tab. 2, we present the results of the BERT-QA model on the NewsVideoQA [2] dataset. We evaluate the model on

Table 1. Performance comparison of BERT-QA model on M4-ViteVQA [3] dataset when the answer to questions is present in the concatenated list of OCR tokens from evenly sampled frames.

Split	Answer present in context	Finetuning	Acc.	ANLS	No. of QA pairs
Task 1 Split 1	No	×	9.03	17.05	1971
Task 1 Split 1	No	✓	21.96	32.18	1971
Task 1 Split 1	Yes	×	19.20	25.25	911
Task 1 Split 1	Yes	✓	47.42	55.14	911
Task 1 Split 2	No	×	8.17	15.81	1321
Task 1 Split 2	No	✓	17.10	26.05	1321
Task 1 Split 2	Yes	×	20.30	27.19	532
Task 1 Split 2	Yes	✓	42.29	48.90	532
Task 2	No	×	10.89	18.41	762
Task 2	No	✓	16.01	24.08	762
Task 2	Yes	×	25.78	30.48	318
Task 2	Yes	✓	38.05	43.82	318

Table 2. In this table, we show the results of the performance of the BERT-QA model on the test set of the NewsVideoQA [2] dataset. For the random frame, we sample a frame randomly and consider its OCR tokens as context to the model.

Training data	Testing data	Ft	Acc.	ANLS
-	single random frame	×	16.78	22.47
single correct frame	single random frame	✓	23.71	29.47
-	single correct frame	×	33.29	43.43
single correct frame	single correct frame	✓	46.55	56.81
-	1fps-sampled-frame	×	31.31	40.60
single correct frame	1fps-sampled-frame	✓	51.25	62.67
1fps-sampled-frame	single random frame	✓	17.41	20.36
1fps-sampled-frame	single correct frame	✓	37.26	42.26
1fps-sampled-frame	1fps-sampled-frame	✓	52.29	61.12

two types of training setups: (a) BERT-QA model trained on single frame OCR information. The single frame is obtained from the timestamp of the video where the question was defined. (b) BERT-QA model trained on the concatenated list of OCR tokens from evenly sampled frames per second. In addition to testing types for a single frame (OCR tokens from the frame at which the question was defined) and concatenated lists of OCR tokens (1fps), we also test by considering OCR tokens from a randomly sampled frame. The results indicate that the model tested on OCR tokens from randomly sampled frames performs poorly. This can be attributed to the nature of the BERT-QA model, which is an extractive QA model that predicts the span of the word it considers as the answer from the context. If the context provided to the model is incorrect or unrelated to the question asked, it will extract an incorrect or irrelevant answer span. However, when the correct frame is provided, i.e., the frame at which the question was defined, the model performs better by obtaining the correct answer as the context is correct. In the `1fps-sampled-frame` experiment, we concatenate OCR tokens from multiple frames (sampled at 1fps), resulting in an increased context. This approach decreases the chances of missing the answer in the context and thus yields improved performance compared to the randomly sampled frames.

In Tab. 3, we present the result of the BERT-QA model's ability to generalize by performing out-of-domain finetuning and testing. We first experiment by testing NewsVideoQA dataset's 1fps test set on BERT-QA trained on M4-ViteVQA dataset's `Task1Split1`. We further finetuning this BERT-QA model trained on M4-ViteVQA `Task1Split1` on the NewsVideoQA dataset (OCR tokens of 1fps). We repeat the same experiments for the second split i.e. `Task1Split2`. It can be seen that the performance by testing a model initially trained on M4-ViteVQA i.e. out-of-domain data, achieves decent performance on a dataset that is completely new or out of domain. By further finetuning this model, it achieves even better performance, thereby leveraging training on the M4-ViteVQA dataset. We also experiment by testing `Task1Split1` on BERT-QA trained on NewsVideoQA. We further finetune it on the M4-ViteVQA dataset. Even in this case, the performance of

Table 3. In this table, we show the results of the performance of the BERT-QA model out of domain data.

Type of training	Continued finetuning	Testing data	Acc.	ANLS
M4-ViteVQA Task 1 Split 1	-	NewsVideoQA	40.39	51.86
M4-ViteVQA Task 1 Split 1	NewsVideoQA	NewsVideoQA	50.41	61.04
M4-ViteVQA Task 1 Split 2	-	NewsVideoQA	37.26	47.36
M4-ViteVQA Task 1 Split 2	NewsVideoQA	NewsVideoQA	52.81	63.54
M4-ViteVQA Task 2	-	NewsVideoQA	34.96	46.98
M4-ViteVQA Task 2	NewsVideoQA	NewsVideoQA	53.44	64.27
NewsVideoQA	-	M4-ViteVQA Task 1 Split 1	7.86	12.68
NewsVideoQA	M4-ViteVQA Task 1 Split 1	M4-ViteVQA Task 1 Split 1	22.17	31.95
NewsVideoQA	-	M4-ViteVQA Task 2	7.34	12.17
NewsVideoQA	M4-ViteVQA Task 2	M4-ViteVQA 2	15.74	24.13

Q. What is the result of her test?

Ground truth : negative
BERT-QA without fine-tuning : **tdy heh negative**
BERT-QA with fine-tuning : **negative**

Q. Who is Bill Gross?

Ground truth : pimco cofounder
BERT-QA without fine-tuning : **worldwide exchange**
BERT-QA with fine-tuning : **pimco cofounder**

Figure 2. Qualitative results from M4-ViteVQA dataset.

the finetuned model after initial training on NewsVideoQA achieves better results than by just finetuning BERT-QA directly on M4-ViteVQA for `Task1Split1`. Therefore it can be inferred that the experiments demonstrate that the BERT-QA model can effectively generalize across domains through out-of-domain finetuning. Additionally, leveraging training on a different dataset improves performance on news datasets showcasing the potential benefits of transfer learning in this task. In conclusion from this section, we can infer that despite the videos being from different domains, it's evident that the model can adapt and assist with the task of extracting text from videos, showcasing its ability to generalize effectively across varied datasets. It not only emphasizes the model's adaptability but also holds the potential to be used in future pretraining techniques, in the context of utilizing combined data from diverse sources, thereby leading to improved overall performance and a deeper understanding of cross-domain information processing.

2. Qualitative results

In this section, we present qualitative analyses conducted on the two datasets which are NewsVideoQA [2] and M4-ViteVQA [3], to gain deeper insights. Fig. 3 showcases qualitative results obtained from the NewsVideoQA dataset. We compared the ground truth with the predictions made by the BERT-QA model before and after finetuning. The results demonstrate that finetuning helps and improves the model's ability to extract relevant answers related to the questions. Similarly for the M4-ViteVQA dataset, we show the qualitative results in Fig. 2. In Fig. 4, we present qualitative results from the NewsVideoQA dataset. We compare the predictions of the BERT-QA models using context from OCR tokens from randomly sampled frames and context from the frame on which the question was defined. The results indicate that text in the random frame is insufficient for the model to obtain accurate answers. However, when provided with OCR tokens from the frame where the ques-

tion was defined, the model successfully obtains the correct answer. In Fig. 5, we show the results for the out-of-domain experiments.

References

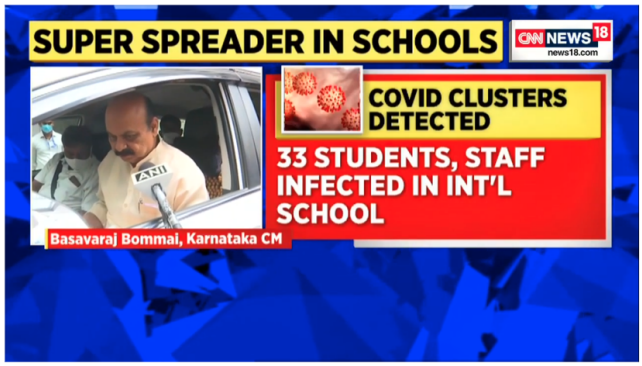
[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019. 1

[2] Soumya Jahagirdar, Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Watching the news: Towards videoqa models that can read. In *WACV*, pages 4430–4439. IEEE, 2023. 1, 2, 3, 5

[3] Minyi Zhao, Bingjia Li, Jie Wang, Wanqing Li, Wenjing Zhou, Lan Zhang, Shijie Xuyang, Zhihang Yu, Xinkun Yu, Guangze Li, Aobotao Dai, and Shuigeng Zhou. Towards video text visual question answering: Benchmark and baseline. In *NeurIPS*, 2022. 1, 2, 3

Q. How was delhi's air on nov 25?

Ground truth	: very poor
BERT-QA without fine-tuning	: **dirty**
BERT-QA with fine-tuning	: **very poor**



Q. In which school where 33 students and staff infected?

Ground truth	: int'l school
BERT-QA without fine-tuning	: **basavaraj bommai**
BERT-QA with fine-tuning	: **int'l school**

Figure 3. This figure shows the qualitative results of NewsVideoQA [2] dataset. We show the ground truth, prediction without finetuning the BERT-QA model, and prediction after or with finetuning the BERT-QA model. It can be seen that finetuning helps the model to extract the answers relevant to the questions.

(correct frame)

(random frame)

Q. Which tour is in jeopardy?

Ground truth	: cricket
BERT-QA with random frame	: **india**
BERT-QA with correct frame	: **cricket**

Figure 4. This figure shows the qualitative results of NewsVideoQA [2] dataset. We show the ground truth, prediction of the BERT-QA model with context as the concatenated list of OCR tokens of the randomly sampled frame, and prediction of the BERT-QA model with context as the concatenated list of OCR tokens of the frame on which the question was defined. It can be seen that the text in the random frame is not sufficient for the model to obtain the answer. Whereas if we give OCR tokens of the frame where the question was defined, it obtains the correct answer.

(M4-ViteVQA dataset)

Q. What is this place named?

Ground truth : great sand dunes
Only M4-ViteVQA : **great sand**
NewsVideoQA + M4-ViteVQA : **great sand dunes**

(NewsVideoQA dataset)

Q. How many mutations did the omicron variant b.1.1529 has in spike protein?

Ground truth : 30+
Only NewsVideoQA : **18**
M4-ViteVQA + NewsVideoQA : **30+**

Figure 5. This figure shows the qualitative results for cross-domain training experiments. It can be seen that, when a dataset is trained on its own training set, and tested, the predictions in some cases are not accurate. Whereas out-of-domain training helps the model to understand the extractive nature of the questions thereby increasing the capability to find the correct answer by providing more context.