SUTD-TrafficQA: A Question Answering Benchmark and an Efficient Network for Video Reasoning over Traffic Events (Supplementary Material)

Li Xu He Huang Jun Liu*
Information Systems Technology and Design
Singapore University of Technology and Design
{li_xu, he_huang}@mymail.sutd.edu.sg, jun_liu@sutd.edu.sg

1. Context-query sub-module

In our context-query sub-module, to facilitate the fusion process between the visual context input and textual query input, the currently selected frame visual feature, I_t, is first projected to the same feature space ($\mathbb{R}^{2d}$) as a single element in textual input ($H^q \in \mathbb{R}^{n_q \times 2d}$, and $H^{a_i} \in \mathbb{R}^{n_{a_i} \times 2d}$). Then following [1], we construct the context-query fusion process as follows: we first compute the similarities between the visual context input (the visual context input contains a single element here, i.e., the frame feature I_t) and each element in textual query input (i.e., each word representation in H^q and H^{a_i}), obtaining two similarity matrices: $S^{I_t,q} \in \mathbb{R}^{1 \times n_q}$ for the frame-question fusion, and $S^{I_t,a_i} \in \mathbb{R}^{1 \times n_{a_i}}$ for the frame-answer fusion. Note the similarity function we use here is the dot product function followed by a $Softmax$ normalization. Then the frame-aware-question representation, $F^{I_t,q}$ is computed as: $F^{I_t,q} = S^{I_t,q} \cdot H^q$, and similarly the frame-aware-answer representation, F^{I_t,a_i}, is computed as: $F^{I_t,a_i} = S^{I_t,a_i} \cdot H^{a_i}$.

The two combined representations, $F^{I_t,q}$ and F^{I_t,a_i}, are then fused with the visual context (I_t):

$$v_t^i = [I_t; F^{I_t,q}; F^{I_t,a_i}; I_t \odot F^{I_t,q}; I_t \odot F^{I_t,a_i}] \quad (1)$$

where $\odot$ represents element-wise product. The concatenated outputs $\{v_t^i\}_{i=1}^N$ (N is the number of candidate answers), denoted as v_t, that incorporates the information of currently selected frame and the QA embedding, can be fed into the interaction LSTM in our Interaction Module for more information interaction to guide the subsequent dynamic reasoning in our Eclipse network.

2. Dataset statistics

As shown in Fig. 1, the wordcloud of questions and candidate answers for our dataset shows that the QA pairs in our dataset mainly focus on various and complex traffic

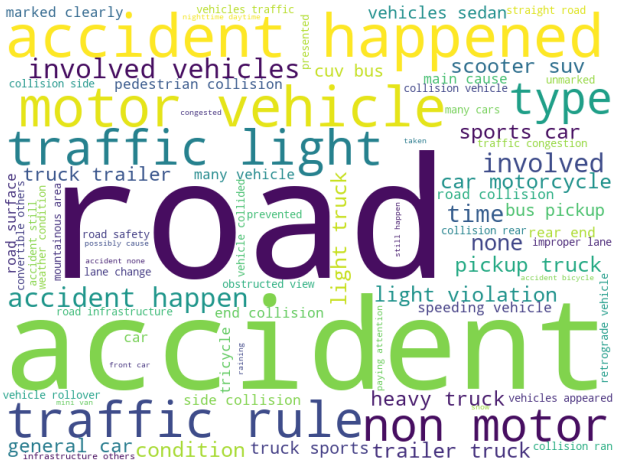


Figure 1. The wordcloud of questions and candidate answers in our SUTD-TrafficQA.

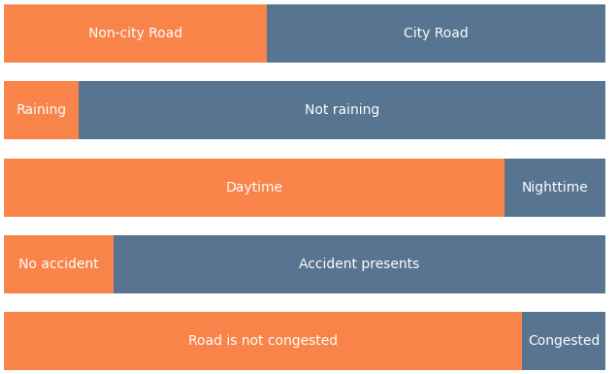


Figure 2. The diversity of videos in terms of different aspects in our SUTD-TrafficQA.

*Corresponding Author.

events. And the Fig. 2 demonstrates the diversity of video collection in our SUTD-TrafficQA in different aspects including road situation, weather, time, and so on. By incorporating the aforementioned diversity in QA pairs and video collection, our dataset shall be able to serve as a comprehensive benchmark for video reasoning of traffic events.

3. Parameter analysis

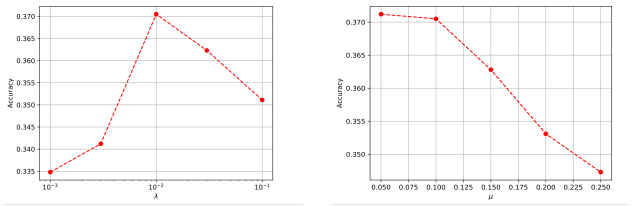


Figure 3. The effects of the parameters, λ and μ, on the reasoning accuracy.

We analyze the effects of the network parameters, λ and μ, on the reasoning accuracy. As shown in Fig. 3, when we increase the λ in our loss function, the accuracy reaches its peak when the λ equals to 0.01. As for the parameter μ, it controls when the network can exit the reasoning process. As shown in Fig. 3, a larger μ means that our network exits reasoning earlier, and thus increasing μ can incur a decrease in reasoning accuracy although with lower computation cost. Considering these experiment results, to maximize the accuracy-to-computation ratio, we set λ to 0.01 and μ to 0.1 in our network to achieve reliable and efficient dynamic reasoning.

4. Qualitative examples

In this section, we present more qualitative examples from our SUTD-TrafficQA dataset to show how our model achieves reliable reasoning in an efficient and dynamic way (recalling that the numbers above the selected frames show the order of the sequence selected by our network).

5. Dataset examples

We present more examples from our dataset as the following. From these examples, we can see that various levels of casual reasoning are required to answer the challenging questions in our dataset, and our dataset covers a wide range of traffic events.

References

[1] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 1

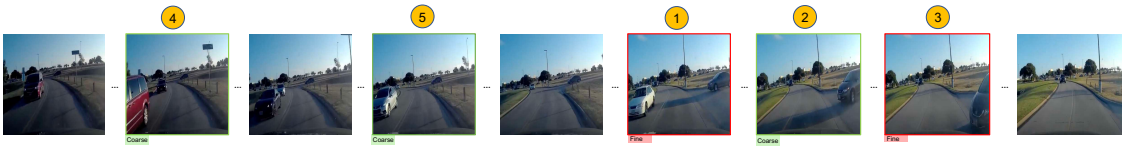
Q: Which car did the white van collide with?

A1: The black car on the left. ✗

A2: The red car on the right. ✗

A3: The black car on the rightmost side. ✓ (Eclipse prediction)

A4: There is no collision event in the video. ✗



Q: Which car did the camera-mounted car collide with?

A1: The red car on the left. ✗

A2: The silver sedan on the left. ✗

A3: The white sedan on the left. ✗

A4: The black sedan on the right. ✓ (Eclipse prediction)

Basic Understanding

Q: What's the condition of the road surface?

✗ The road is dusty and muddy.

✗ The road is covered by snow and ice.

✓ The road is smooth and clean.

✗ The road is very uneven.

Attribution

Q: What might be the reason that led to the rollover of the SUV?

✗ Another car hit the SUV.

✗ The SUV crashed into a tree.

✗ The road is slippery.

✓ The SUV crashed into another car.

Introspection

Q: Is there anything could have been done to avoid this accident?

✗ The accident could have been avoided if there had been no pedestrian.

✓ The accident could have been avoided if the moving SUV had slowed down.

✗ The accident could have been avoided if the road had been marked clearly.

✗ The accident could have been avoided if the road had been not slippery.

Counterfactual Inference

Q: Would the accident still happen if the road were wider?

✓ Yes, a wider road does not help the situation.

✗ No, a wider road would have provided enough space to safely avoid the accident.

✗ No, there is no accident.

Event Forecasting

Q: What will happen to the moving SUV soon?
(Showing model the video up to 00:04 only.)

✓ The moving SUV will roll over.

✗ The moving SUV will continue to move forward.

✗ The moving SUV will turn left.

✗ The moving SUV will crash into a tree.

Reverse Reasoning

Q: What might happened moments ago?
(Showing model the video from 00:04 onwards only.)

✗ The SUV was hit by another vehicle from the back.

✗ The wet road caused the SUV losing its control.

✓ The SUV crashed into the parked car.

✗ The tree fell down.

Basic Understanding

Q: How many vehicles rolled over?

- ✗ None.
- ✓ One.
- ✗ Two.
- ✗ Three.

Attribution

Q: What might be the reason which led to this accident?

- ✗ Improper lane changing.
- ✓ Violation of traffic lights.
- ✗ Extreme weather condition.
- ✗ The road is not marked clearly.

Introspection

Q: What could have been done to prevent this accident from happening?

- ✗ The accident could have been avoided if the white SUV had stayed on its own lane.
- ✗ The accident could have been avoided if the yellow school bus had turned right.
- ✗ The accident could have been prevented if the white sedan had turned left.
- ✓ The accident could have been avoided if the black sedan had obeyed the traffic light.

Counterfactual Inference

Q: Would the accident still happen if there were no yellow school bus?

- ✓ Yes, the yellow school bus has nothing to do with this accident.
- ✗ No, the accident would not happen since the yellow school bus obstructed the road.
- ✗ No, the accident would not happen since the yellow school bus crashed into the white SUV.

Event Forecasting

Q: Which event will happen next?

(Showing model the video up to 00:05 only.)

- ✗ The black sedan will turn left.
- ✗ The yellow school bus will pass the intersection.
- ✗ The white sedan will move forward.
- ✓ The white SUV will roll over.

Reverse Reasoning

Q: What might happened before?

(Showing model the video from 00:04 onwards only.)

- ✗ The white SUV lost its control.
- ✗ The yellow school bus crashed into the white SUV.
- ✓ The black sedan violated the traffic light and crashed.
- ✗ The black sedan turned left.

Basic Understanding

Q: How much damage does the black sedan receive after the accident?

- ✗ There is no visible damage.
- ✗ There is no accident in the video.
- ✗ Only minor scratches.
- ✓ The black sedan is severely damaged.

Attribution

Q: Which car is responsible for the accident?

- ✗ The blue SUV.
- ✓ The white SUV.
- ✗ The black sedan.
- ✗ The white sedan.

Introspection

Q: What could have been done to avoid this accident?

- ✗ The accident could have been avoided if there had been safety barriers.
- ✗ The accident could have been avoided if the white SUV had slowed down.
- ✗ The accident could have been avoided if the road had been marked clearly.
- ✓ The accident could have been avoided if the white SUV had not crossed the lane.

Counterfactual Inference

Q: Would the accident still happen if the white SUV stayed on its own lane?

- ✗ Yes, the accident would still happen.
- ✓ No, the accident would not happen if the white SUV stayed on its own lane.
- ✗ No, there is no accident.

Event Forecasting

Q: Will any accidents happen soon?

(Showing model the video up to 00:06 only.)

- ✓ Yes, the white SUV will crash into some car in the opposite way.
- ✗ No, the vehicles are running normally on the road.
- ✗ Yes, the white SUV will crash into the pedestrian.

Reverse Reasoning

Q: What might happened before?

(Showing model the video from 00:07 onwards only.)

- ✗ The black sedan turned right.
- ✗ The white SUV was hit by another car.
- ✓ The white SUV crossed the center lane.
- ✗ The white sedan lost its control.