Supplementary Material for TDIUC-AVQA: A Visual Question Answering Dataset in Low-Resource Assamese Language

Nazreena Rahman[1[0000-0001-7188-0111]], Pankaj Choudhury[2[0009-0001-1159-3118]], Prithwijit Guha[1,2[0000-0003-2885-0026]], Ashish Anand[2,3[0000-0002-0024-3358]], and Sukumar Nandi[2,3[0000-0002-5869-1057]]

[1] Department of Electronics and Electrical Engineering
[2] Centre for Linguistic Science and Technology
[3] Department of Computer Science and Engineering
Indian Institute of Technology Guwahati, Assam, India
{nazreena, pankajchoudhury, pguha, anand.ashish, sukumar}@iitg.ac.in

1 Results and Discussion

1.1 Quantitative Analysis

The initial experiment focused on training the proposed models on the TDIUC-AVQA dataset, with the RNN layer size increased up to 5. As shown in Table 1, the test results indicate that Bi-GRU achieved the highest precision and recall scores for layer 1. For layer 2, BiGRU obtained the highest F1-score and accuracy values when using soft attention. The evaluation results confirm that the BiGRU model outperforms other models as a question encoder.

Table 2 shows a decrease in category-wise performance when the proposed AVQA model is trained without *Absurd* category. This suggests that including the *Absurd* category in training helps mitigate language prior bias, thereby enhancing model performance.

1.2 Qualitative Analysis

Fig. 1 shows the examples of qualitative results with better visibility.

Table 1: Performance comparison of the proposed AVQA method using various question encoders across different RNN layers with Soft Attention on the TDIUC-AVQA dataset

Question Encoder	No. of RNN Layers	Precision	Recall	F1	Accuracy
LSTM	1	79.17	75.96	76.4	75.75
	2	80.41	77.72	77.96	77.6
	3	80.25	77.93	78.06	77.81
	4	80.59	78.13	78.35	77.99
	5	80.24	77.66	77.97	77.52
Bi-LSTM	1	80.88	79.88	79.87	79.73
	2	80.09	77.50	77.77	77.34
	3	80.66	79.12	79.38	79.01
	4	80.83	78.68	78.91	78.18
	5	80.46	78.07	78.62	77.96
GRU	1	80.14	79.27	79.22	79.13
	2	79.74	77.94	78.2	77.79
	3	80.57	77.89	77.99	77.73
	4	80.22	77.17	77.66	77.06
	5	80.14	77.45	77.75	77.30
Bi-GRU	1	80.91	**80.27**	79.99	**80.1**
	2	**81.22**	80.21	**80.31**	80.07
	3	80.48	77.78	78.15	77.64
	4	80.91	78.8	79.13	78.69
	5	80.85	78.65	79.01	78.53
Transformer Encoder	1	76.74	76.42	76.03	76.16
	2	77.44	78.02	77.77	77.25
	3	76.21	77.49	76.42	77.26
	4	76.48	77.58	76.68	77.37
	5	20.05	24.98	20.71	24.25

Table 2: Performance evaluation of the proposed AVQA method on data excluding samples from the *Absurd* category during training

Question Category Wise	Precision	Recall	F1	Accuracy
Object Presence	91.27	91.27	91.27	91.27
Subordinate Object Recognition	82.31	82.82	81.61	81.32
Counting	44.75	44.56	38.66	44.55
Color Attributes	47.83	45.48	44.14	45.29
Other Attributes	41.73	45.09	40.76	40.24
Activity Recognition	53.98	48.01	46.96	47.73
Sport Recognition	**93.04**	**92.32**	**92.58**	**92.18**
Positional Reasoning	23.65	29.18	22.54	24.09
Scene Classification	73.34	55.60	63.25	60.17
Sentiment Understanding	53.42	51.72	51.96	47.48
Utility/Affordance	63.53	32.00	33.85	14.04

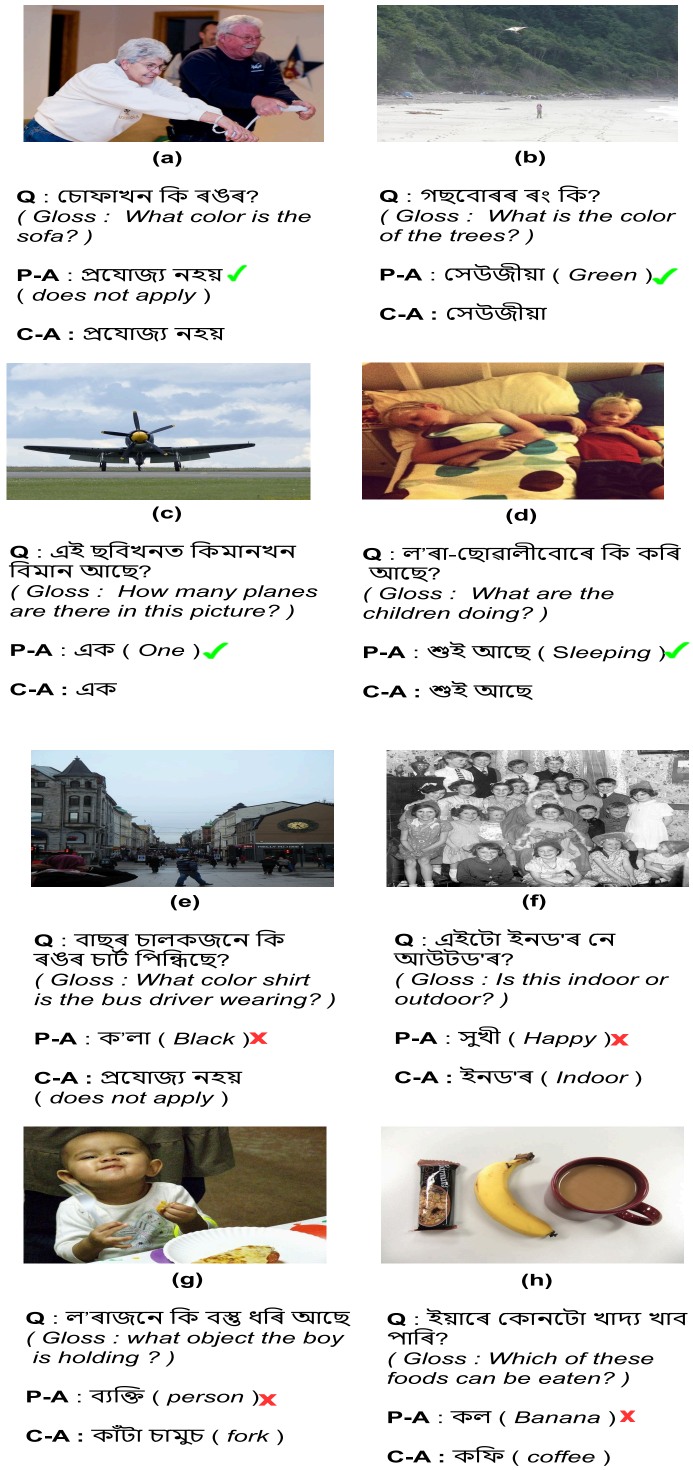

Fig. 1: Examples of qualitative results for the proposed AVQA Method by using Bi-GRU as an Question Encoder. The first two rows display accurately classified answers, while the third and fourth rows highlight the inaccurately classified answers. Q – Question associated with the image, P-A – Predicted answer, C-A – Correct answer, and Gloss – refers to gloss annotation.