International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

https://ijcjournal.org/index.php/InternationalJournalOfComputer/index

# Vehicle Feature VQA: Visual Question Answering for Vehicle Feature

Pa Pa Tun[a]*, Khin Mar Soe[b]

*[a,b]Faculty of Computer Science, University of Computer Studies Yangon, Yangon, Myanmar*

*[a]Email:imageproject2025@gmail.com*

*[b]Email:khinmarsoe@ucsy.edu.mm*

**Abstract**

Visual Question Answering (VQA) can automatically produce the predict answers for questions and real-world images. In this paper, we propose the VQA dataset for Vehicle Feature to know the knowledge of Vehicle. We develop the VQA model using RestNet50 in Convolutional Neural Networks (CNN) for feature extraction of images and Long Short-Term Memory (LSTM) for question feature extraction and answer generation. The experimental result describes the training loss, evaluation loss, Blue Score, and VQA accuracy for epochs 20 and epochs 30. In epochs 20, after VQA model generated the training loss 1.1949 , evaluation loss 1.7953, Blue Score 0.6180, and VQA accuracy 0.0493, this model predicted the one correct answer for question and image. In epochs 30, the VQA model predicted the five correct answers in fifteen test data for vehicle feature questions and image according to generate the training loss 0.8780, evaluation loss 1.6634, Blue Score 0.6775 and VQA accuracy 0.0627.

***Keywords:*** Visual Question Answering; Vehicle Feature; RestNet50; Convolutional Neural Networks; feature extraction; Long Short-Term Memory; Blue Score.

------------------------------------------------------------------------

*Received: 12/10/2025*
*Accepted: 2/10/2026*
*Published: 2/21/2026*

------------------------------------------------------------------------

* *Corresponding author.*

## 1. Introduction

In an education sector, Visual Question Answering can be supported visual learning to a great extent and used for Visual Chart bots for Education, Gamification of VQA Systems, and Automated Museum Guides [1,2]. Visual Question Answering (VQA) is a vision-language task to generate the answer as output according to image and text-based question input as a human-like manner [2,3]. VQA retrieves automatically an answer for asking question about image using natural language processing and computer vision [3,4,5].

Several Visual Question Answering are created to obtain the VQA models using Convolutional Neural Network (CNN) such as VGGet for image feature extraction. VQA are commonly used Recurrent Neural Network (RNN) for question feature extraction. In VQA, RNNs have problems to explode and vanish gradients during training a deep neural network [2,6,7]. Visual Question Answering system consists of image feature extraction, question feature extraction, feature conjugation, and answer generation [8,9,10]. The popular VQA datasets are COCO-QA Dataset, CLEVR, DAQUAR, and Visual7W dataset [7,11,12].

In this paper, the 532 vehicle images and 195 KB train data and 124KB evaluate data are collected for creating VQA dataset for vehicle's features in this research. Firstly, the question and response data are built to obtain the vocab questions and vocab answers using tokenizer. Then, answer train data, image and transform are put into the VQA dataset. Image is pretrained into generate feature extraction using ResNet50 in Convolutional Neural Networks (CNNs). Long Short-Term Memory (LSTM) are used to generate Question encoder and Answer decoder. Then, VQA model are created and questions vocab size and answer vocab size are put in VQA model. VQA model are trained by putting the training loader, and evaluation loader into the VQA model. The output of training model is produced training loss, evaluation loss, accuracy of VQA and blue score. Finally, the result of prediction VQA model is produced.

## 2. Vehicle Feature for VQA Dataset

Vehicle Feature dataset is developed into VQA dataset in 2025. This is purposed to research and education purpose for vehicle feature knowledge. This dataset includes 532 vehicle images and 71 KB evaluation data and 117KB training data. This dataset also provides questions for asking vehicle's feature knowledge and provides correct answers for vehicle's feature as shown in Figure 1.

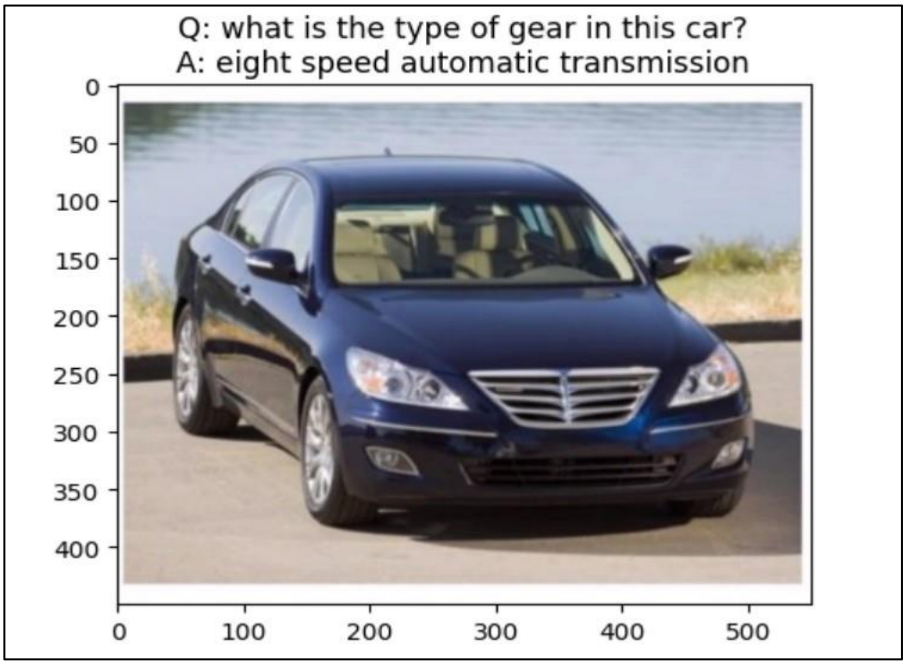

Q: what is the type of gear in this car?
A: eight speed automatic transmission
0
50
100
150
200
250
300
350
400
0
100
200
300
400
500

**Figure 1:** Answer of vehicle's feature question

This dataset also has responses for presenting fully sentences answering of questions. This dataset includes 195 KB train data and 124 KB evaluate data for collecting questions and answers and responses. The full_answer-train data are created into data frame. This dataset collects questions and response pair from dataframe to collect all_question_answer_pair data. In Figure 2 shows the question and response for answering question.

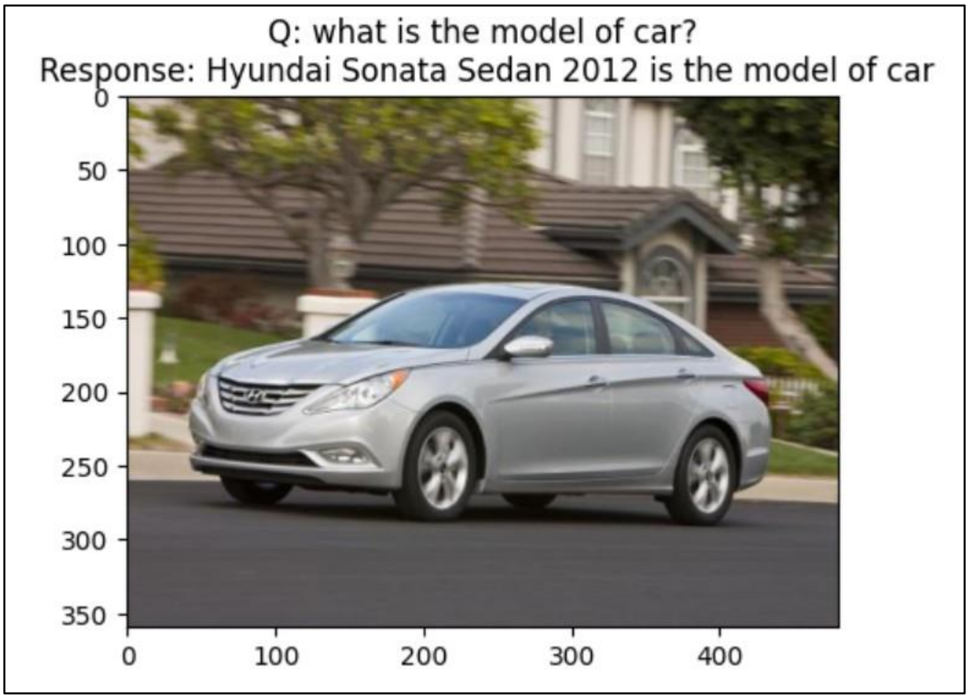

Q: what is the model of car?
Response: Hyundai Sonata Sedan 2012 is the model of car
0
50
100
150
200
250
300
350
0
100
200
300
400

**Figure 2:** Response for answering question

The 13750 words in the vehicle feature questions are used to generate a bag-of-words representation. The bad-of-words model is first built vocabulary that can be word-based as words in texts. The vocabulary contains the token unknown represents as zero, token padding represents as one, start of sentence represents as two, and end of sentence represents as three. The index lets four. The first word of sentence can be represented as index are built into vocabulary. Then, the index is increased one. The other words of sentence are represented as index. Then, the questions and responses are built into vocabulary. The questions vocabulary size of length and answers vocabulary size length is defined. In the answer vocabulary, the index is represented key and the word is represented value. In the question vocabulary, the index is represented key and the word is represented value. The maximum length of question is defined and the maximum length of response is defined.

VQA dataset is created by using image folder and dataset. The row of dataset is read. In the image folder, the path of image is recorded. The row of question text is converted into a numerical representation is called embedding of question. The row of response text is converted into a numerical representation is called embedding of response. The image is converted into RGB color. The image is transformed to resize the image as (224 ,224) size and to change the random rotation and to convert the image into numerical representation and to normalize. The train VQA dataset is created and train loader is created. The evaluation VQA dataset is created and evaluation loader is created.

## 3. Methods

In this section, we propose image feature extraction, question encoder, and answer decoder to develop VQA model for vehicle feature. We use Convolution Neural Networks (CNN) for image feature extraction and Long Short-Term Memory (LSTM) for question encoder and answer decoder. We describe the three steps for VQA model the following phases:

### *3.1. Image Feature Extraction*

In VQA model, feature extraction converts vehicles image data into relevant features to classify the images based on patterns, textures, colors, and structures within the image. The image feature extraction uses pre-trained models such as ResNet50 is a deep convolutional neural network (CNN). Vehicle images feed into pre trained ResNet50 model after they are resized to 224*224 pixels and normalized. ResNet50 performs leveraging the pre-trained weights from models trained on vehicle datasets. ResNet50 passes an image through its layers to generate a high-dimensional feature vector such as 2048 dimensional.

### *3.2. Question Encoder*

In VQA model, Question Encoder processes encoding of the vehicle feature questions to extract relevant entities and concepts. Question Encoder uses Long-Short-Term Memory (LSTM) that obtain input on the question vocabulary size of length as 169 and embedding dimensions as 256. The LSTM with two hidden layers is used to obtain 2048-dim embedding for the question because of performing last hidden state representations as 512 dimensions from each of the two hidden layers (2*2*512). LSTM takes sequence of words embeddings and processes them to produce final representation.

### 3.3. Answer Decoder

Answer decoder executes to create the final answer based on decoding of the vehicle feature answers in the VQA model. It embeds the 256 embedding size and 940 answer vocabulary size to produce the embedding result. It uses three LSTM layers, respectively, with input of 512 hidden dimension and 0.2 dropout and adding embedding size with1024. It carries the transformer attention with the input of 512 hidden dimensions and linear input of hidden dimensions and answer vocabulary size to calculate the weight for making the next output word. It receives an integrated representation of the vehicle feature image and question. It uses the beam search algorithm to enable for non-greedy local decisions that can expedite a sequence with a higher overall probability. It generates a preferred candidate answer based on the encoded image and question.

## 4. Experimental Results

For the experiment, we processed two steps that are training VQA and testing VQA methods. The training 195 KB data put in train loader and the evaluation 124 KB data turn over into evaluation loader according to 532 images. The training VQA method is performed as train VQA model with the input of train loader, evaluation loader, number of epochs, criterion, and optimizer and then predicts answers according to perform tensor to text of images and questions. This method saves the path of VQA model. The training VQA method evaluates the VQA model accuracy, training loss, evaluation loss, BLUE Score and time for each epoch. In Figure 3 illustrates the training loss and evaluation loss according to put epoch 30.

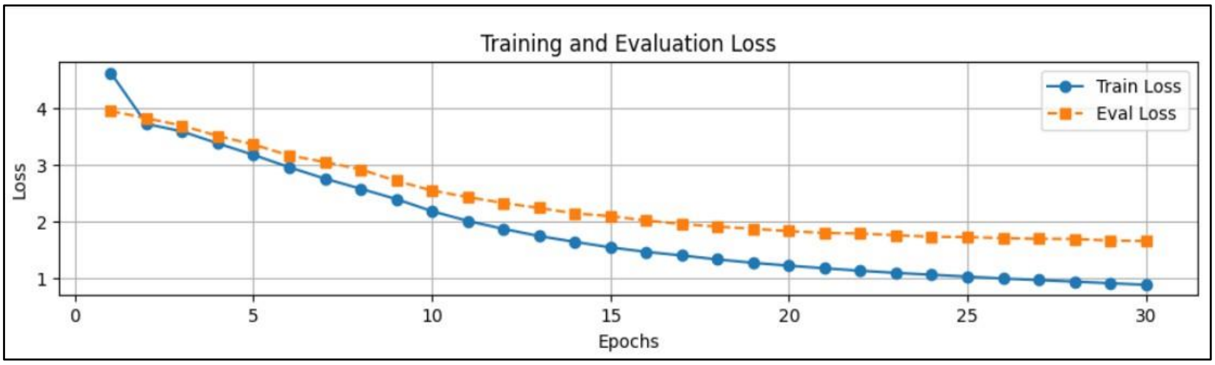


**Figure 3:** Training and evaluation loss

In Figure 3 shows the training loss and evaluation loss decrease with producing the train loss 0.8942 and the eval loss 1.6639 at the epochs 30. So, this model is good learning and generalization because of the training loss and the evaluation loss decrease. The Blue Score method contains the input of predicted answer test and answer test to evaluate the Blue Score value. Firstly, this method evaluates precision result by using ngram method. Then, this method calculates the geometric mean values with the input as precision result. The brevity_penalty method uses the input of predicted answer test and answer test to produce the brevity_penalty value. The Blue Score method evaluates the multiplications of brevity_penalty value and geometric value to produce the Blue Score result. If Blue Score is higher, the model is better quality for human level quality. In Figure 4 displays the Blue Score result 0.6455 with the epochs 30 to achieve the good quality of model.

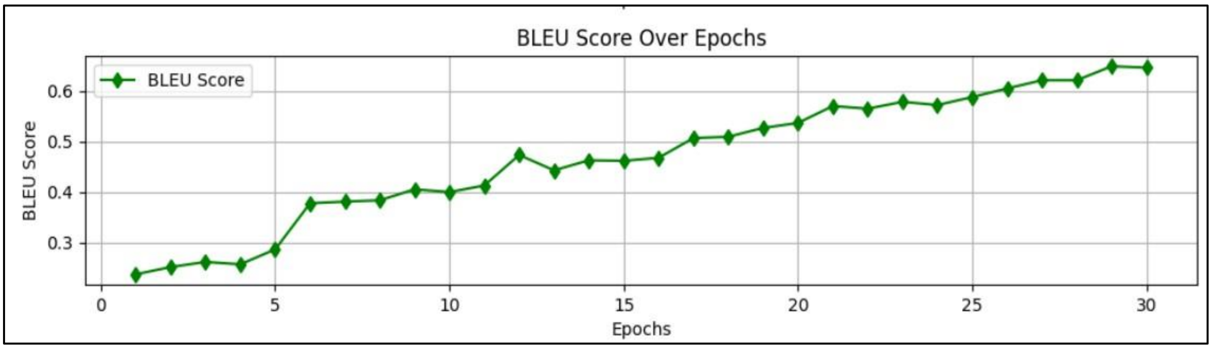


**Figure 4:** Blue score result

The VQA_ accuracy method calculates the VQA accuracy score for the input of predicted answer and ground truth answer. This method normalizes the predicted answer and ground truth answers to change the lowercase for all characters and to remove articles and periods and to convert number words to digits and to replace other punctuation with spaces. If normalized predicted answer is equal to the normalized ground truth answer, then match_count adds one. Finally, this method calculates the multiplications of 0.3 and match_count and then nominate minimum number between the multiplications result and one. We calculate the average VQA accuracy. We use epochs 30 to obtain the VQA accuracy 0.0493 as shown in Figure 5.

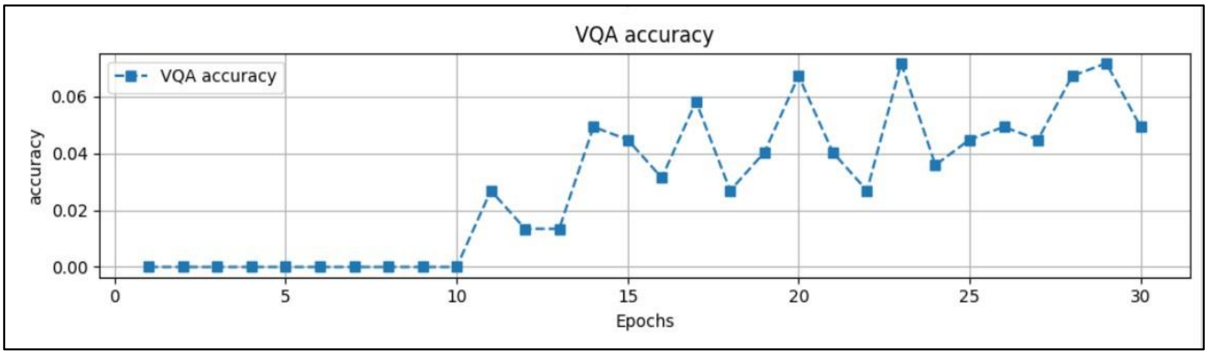


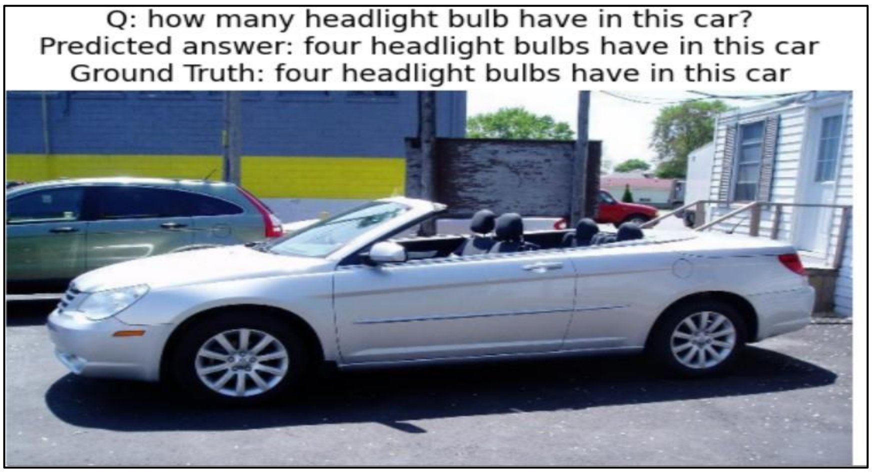
**Figure 5:** VQA accuracy

We compare the results of epochs 20 and epochs 30 according to training loss, evaluation loss, Blue Score and VQA accuracy to show in Table 1. In epochs 20, this model is training five times to produce five VQA accuracy that compare to select the best result.

**Table 1:** VQA performance with epochs 20 and 30

| Epochs | Training loss | Evaluation loss | Blue Score | VQA accuracy | Test data | Correct output | Similar or not correct |
|---|---|---|---|---|---|---|---|
| 20 | 1.1949 | 1.7953 | 0.6180 | 0.0493 | 15 | 1 | 14 |
| 20 | 1.2250 | 1.8449 | 0.5782 | 0.0448 | 15 | 1 | 14 |
| 20 | 1.2733 | 1.8473 | 0.5314 | 0.0731 | 15 | 1 | 14 |
| 20 | 1.2902 | 1.9155 | 0.5074 | 0.0358 | 15 | 1 | 14 |
| 20 | 1.2591 | 1.8616 | 0.5165 | 0.0806 | 15 | 0 | 15 |
| 30 | 0.9887 | 1.6850 | 0.6632 | 0.0448 | 15 | 2 | 13 |
| 30 | 0.8942 | 1.6639 | 0.6455 | 0.0493 | 15 | 2 | 13 |
| 30 | 0.9322 | 1.7439 | 0.6535 | 0.0672 | 15 | 2 | 13 |
| 30 | 0.9145 | 1.6843 | 0.6639 | 0.0582 | 15 | 3 | 12 |
| 30 | 0.8780 | 1.6634 | 0.6775 | 0.0627 | 15 | 5 | 10 |

In epochs 20, the VQA model have training loss 1.1949 smaller than 1.2250 and have evaluation loss 1.7953 smaller than 1.8449 and have Blue Score 0.6180 larger than 0.5782 and then obtain VQA accuracy 0.0493 larger than 0.0448 so that the test data 15 are analyzing this model to obtain one correct output and fourteen not correct output. The VQA accuracy 0.0493 is larger than the other VQA accuracy at epochs 20 and then the training loss 1.1949 and the evaluation loss 1.7953 are smaller than other training loss and evaluation loss.If training loss and evaluation loss are smaller, then Blue Score and VQA accuracy are larger for this model at epochs 20. In epochs 20, this model produces the correct predicted answer according to VQA accuracy 0.0493 as shown in Figure 6.



Q: how many headlight bulb have in this car?
Predicted answer: four headlight bulbs have in this car
Ground Truth: four headlight bulbs have in this car

**Figure 6:** Correct predicted answer for epochs 20

In epochs 30, the VQA model generates the training loss 0.9887 is larger than 0.8780 training loss and other training losses, evaluation loss 1.6850 is larger than 1.6634 evaluation loss and other evaluation loss, Blue Score 0.6632 is smaller than 0.6775 and other Blue Score, and VQA accuracy 0.0448 is smaller than 0.0627 and other VQA accuracy and then this model predicts the two accurate answers and thirteen inaccurate answers for questions and images. So, VQA accuracy 0.0627 and Blue Score 0.6775 have five correct answers and ten inaccurate answers in fifteen test data. In epochs 30, the VQA model produces training loss 0.8780, evaluation loss 1.6634, Blue Score 0.6775, and VQA accuracy 0.0627 and then this model predicts the five correct answers for vehicle questions and answers as presented in Figure 7.

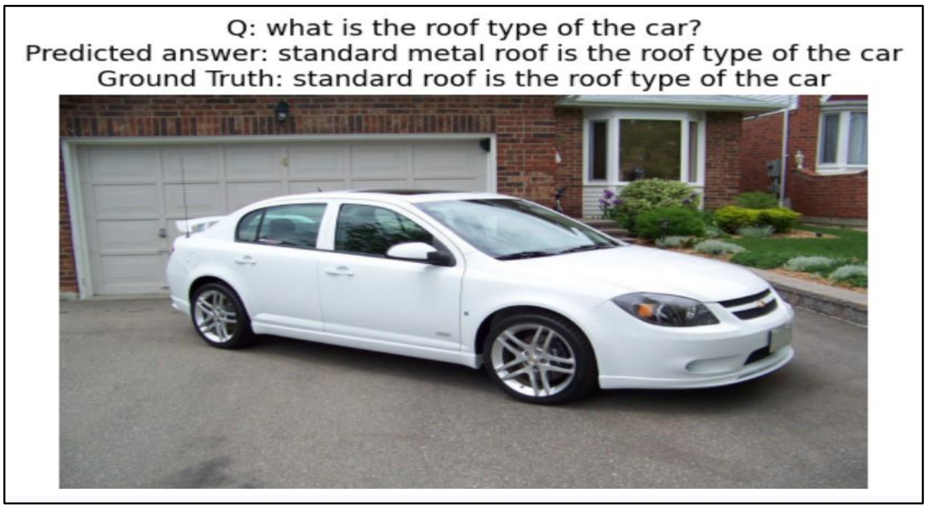

Q: what is the roof type of the car?
Predicted answer: standard metal roof is the roof type of the car
Ground Truth: standard roof is the roof type of the car

**Figure 7:** Correct predict answer for epochs 30

## 5. Conclusion

In this paper, we propose the VQA dataset for Vehicle feature and VQA model with feature extraction of images using ResNet50 in Convolutional Neural Networks (CNN) and questions and answers decoder using Long Short-Term Memory (LSTM). VQA model performance analyze the training loss, evaluation loss, Blue Score and VQA accuracy according to epochs 20 and epochs 30. If the training loss and evaluation loss are large, the model predicts inaccurate answer for vehicle feature in images and questions. If VQA accuracy are large, the model predicts correct answer for images and questions. In future, we must create VQA dataset for Vehicle features and VQA model using Myanmar language.

## Acknowledgement

We would like to grateful for CYIZHUO researcher because of he uploads Stanford cars dataset in GitHub. We pick up images from his Stanford cars dataset.

## References

[1] S. Chowdhury, B. Soni, “eaVQA: An Experimental Analysis on Visual Question Answering Models”, in *Proc. of the 18th International Conference on Natural Language Processing,* 2021, pp. 550-554.

[2] Z.Wang, S. Ji, "Learning Convolutional Text Representations for Visual Question Answering", in *Proc. the 2018 SIAM International Conference on Data Mining,* 2018, pp. 594-602.

[3] S. Antol *et al*., "VQA: Visual Question Answering," in *Proc. 2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 2425-2433.

[4] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra and D. Parikh, "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering," in *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 6325-6334.

[5] J. Guo *et al*., "From Images to Textual Prompts: Zero-shot Visual Question Answering with Frozen Large Language Models," in *Proc. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 2023, pp. 10867-10877.

[6] E. Borisova, N. Rauscher, G. Rehm, "SciVQA 2025: Overview of the First Scientific Visual Question Answering Shared Task", in *Proc. of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, Vienna, Austria, 2025, pp. 182-210.

[7] C. Zhou, G. Chen, X. Bai, M. Dong, "On the Human-level Performance of Visual Question Answering", in *Proc. of the 31st International Conference on Computational Linguistics*, Abu Dhabi, UAE, 2025, pp. 4109-4113.

[8] Z. Zhang, "Enhanced Textual Feature Extraction for Visual Question Answering: A Simple Convolution Approach", *arXiv:2405.00479v2*[online], pp.1-12, https://arxiv.org/html/2405.00479v2 [11.Nov.2024].

[9] Huynh, N.D., Bouadjenek, M.R., Aryal, S., Razzak, I. and Hacid, H. "Visual question answering: from early developments to recent advances--a survey". *arXiv preprint arXiv:2501.03939[on-line],* https://arxiv.org/abs/2501.03939, [11.Jan.2025].

[10] S. Gautam, V. Thambawita, M. Riegler, P. Halvorsen, S. Hicks, "Medico 2025: Visual Question Answering for Gastrointestinal Imaging", *arXiv preprint arXiv:2508. 10869 [on-line],* https://arxiv.org/abs/2508.10869 [14.Aug.2025].

[11] I. Allaouzi, M. B. Ahmed, B. Benamrou, "An Encoder-Decoder model for visual question answering in the medical domain", *CEUR-WS.org*[on-line], vol. 2380, pp.124-132, https://ceur-ws.org/Vol-2380/paper_124.pdf [9-12 September 2019].

[12] R. Pal, S. Kar, D. K. Prasad, "NorVivqA: Visual Question Answering for Visually Impaired in Norwegian Language", *CEUR-WS.org* [on-line], Vol.3975, pp.3-13, https://ceur-ws.org/Vol-3975/paper3.pdf [17-18, June 2025].