



Article Title: Development of Counting Based Visual Question Answering System by Using Transformer and Pyramid Networks with Hybrid Deep Learning Model

Development of Counting Based Visual Question Answering System by Using Transformer and Pyramid Networks with Hybrid Deep Learning Model

S. Nigisha¹, K. Anugirba²

¹ Computer Science Engineering, Bethlahem Institute of Engineering, Karungal, Anna University.

² Computer Science Engineering, Faculty, Bethlahem Institute of Engineering, Karungal, Anna University.

ABSTRACT

Visual Question Answering (VQA) merges images and natural language processing, that enables machines to respond to queries about visual content with prowess by comprehending visual features and contextual cues in text. VQA aims to bridge the gap between human-like understanding and visual comprehension. Counting-based VQA is a specific subfield within VQA that focuses on answering questions related to counting objects or quantities in images. The objective of counting-based VQA is to develop algorithms and models capable of accurately answering questions that involve counting specific objects or quantities in visual data. Our Model consists of Bidirectional Encoder Representations from Transformers (BERT) to extract the texture features from the Question part and for the visual part, Feature Pyramid Network (FPN) is used to extract the deep features from images. Both the textual and visual features are integrated to form a combined set of features. These fused features are fed in to a hybrid model for answer prediction. This hybrid model is an integration of Gated Recurrent Unit (GRU) and One-Dimensional Convolutional Neural Network (1DCNN).

Keywords: Counting based visual question answering (VQA), Bidirectional Encoder Representations from Transformers (BERT), Feature Pyramid Network (FPN), Deep Features, Gated Recurrent Unit (GRU), One-Dimensional Convolutional Neural Network (1DCNN).

1 Introduction

To predict the answer conditioned on the input image and question as a classification task instead of a sentence generation task. Specifically, the final answer can be selected from the class with the highest probability [1]. Fusion of these features to obtain a single feature vector representing both the visual information and the question. The prediction is based on this vector [2]. To encode the textual question Q containing M words using a pretrained BERT. This results into a sequence q of M d -dimensional vectors [3]. The region features are good at capturing objects in an image, and thus very useful for answering questions on object counting [4]. The model attends over the regions in each image separately using the question embedding. Pooling the region features gives a representation of an image. These are concatenated and



Article Title: Development of Counting Based Visual Question Answering System by Using Transformer and Pyramid Networks with Hybrid Deep Learning Model

combined by the question embedding to give the joint scene representation [5]. The two final concatenated vectors and also considers only visual feature vectors and word embeddings. This way, which feature set gives the best accuracy can be observed [6].

2 Recent Works

When the number of decoding layers is increased from one layer to two layers, the impact of the model is greatly enhanced [7]. The multi-scale information of images can be captured by using the high resolution of low-level features and rich semantic information of high-level features. FPN extracts the features from the deep-layer network and gets the same features as the shallow layer features through the up-sampling. Then, the fusion of the two features is used as the input of the classification network and achieves good results in the classification task [16]. Both features are fused by element-wise multiplication and fed into multiple layers of network model to predict the answer [8]. Modified versions from Recurrent Neural Network (RNN) are utilized to better performance such as long short-term memory (LSTM) and GRU. GRU is a modified version from LSTM with 2 gates instead of 4 as in LSTM, so GRU needs less computational power than LSTM [9]. For sufficiently complex spaces, the agent needs to hold this information in memory for a long time. This motivates the need for an explicit external memory representation that is filled by the agent as it interacts with its environment. This memory must be both spatial and semantic so it can represent what is where. We propose a new recurrent layer formulation: Egocentric Spatial GRU (esGRU) to represent this memory [14]. VQA requires a high level understanding of images and questions, and is often considered to be a good proxy for visual reasoning. However, it is not straightforward to use ConvNets in a context where a high level of reasoning is required [10]. VQA challenge, linked to the VQA task, combine joint embedding with some type of attention model to focus on a specific part of the image [11].

The Prediction is done by multiplying output probabilities of both models for each class and taking the answer with the highest value as output [12]. The hybrid model can generate accurate results when compared to a single model. The convolutional processing uses multiple convolution kernels to extract the various features that contribute to classification [13]. The CNN is used to extract spatial features while the GRU is used to extract temporal features. Single-task CNNs are suitable for spatial feature extraction and single-task GRUs are suitable for temporal feature extraction [15]. The semantic features extracted from the question as guiding features for the attention weight calculation of image features so as to use the key information in the question to improve the accuracy of finding the answer from the picture. [17]. The main objective of our work is to develop the VQA model that should take less amount of training time, should be able to accurately answer for simple and complex questions as well. Proposed Work Explanation



Article Title: Development of Counting Based Visual Question Answering System by Using Transformer and Pyramid Networks with Hybrid Deep Learning Model

3.1 Dataset Description

Our proposed method involves gathering visual question and image data from <https://visualqa.org/download.html/>. This dataset comprises open-ended questions related to images and consists of 50,000 images and 150,000 different question types. The corresponding questions and answers are stored in .json file. The input questions follow a data structure defined by task type, data type and data subtype. Similarly the answer structure includes data type, data subtype, answer type, multiple-choice answer and answer confidence.

To implement our method, we specifically utilize the data type “how many” since it suits our counting-based system. The collected images are genuine and balanced, devoid of abstract scenes. Our approach considers three distinct question types per image, which encompass both simple and complex variations. The input images and questions are denoted as I_{input} and Q_{input} . Figure 1 contains sample images from vqa dataset.

Sample Images	Questions	Answers
	How many boys sitting?	3
	How many oranges have a sticker?	2



Article Title: Development of Counting Based Visual Question Answering System by Using Transformer and Pyramid Networks with Hybrid Deep Learning Model

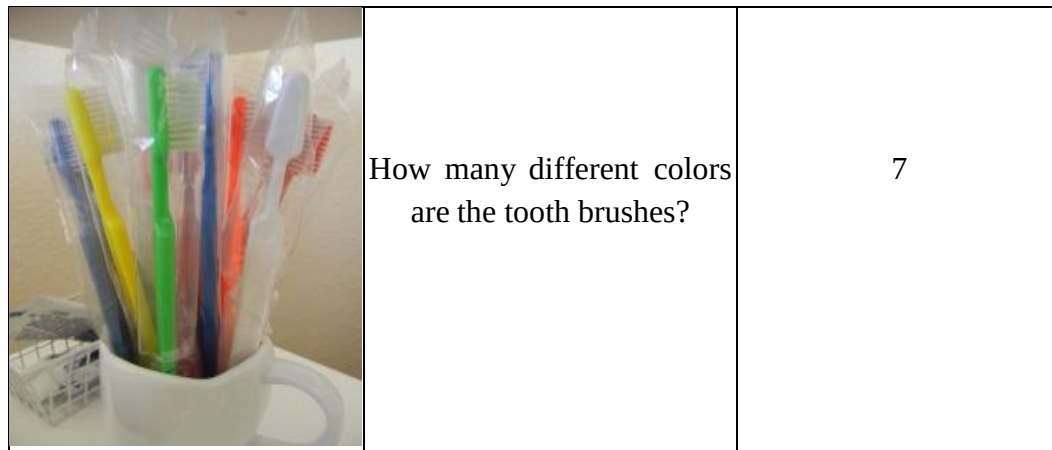


Figure 1: Sample images and questions of the proposed counting based VQA model

3.2 Text Features

Preprocessing refers to the procedure of removing unnecessary data from a given dataset and preparing it for subsequent analysis. In this context, preprocessing encompasses the elimination of punctuation and the application of stemming techniques. By removing punctuation marks and applying stemming the text processing approach achieves improved performance and reduced memory usage. Consequently, the preprocessed data denoted as Qaprep, is obtained without punctuation marks and with applied stemming from Qainput.

To convert text into vectors using BERT, the Qaprep is tokenized into individual words or subwords. Each token is then mapped to its corresponding word embedding vector using a pretrained BERT model. These word embeddings not only represent the meaning of individual words. But also, the context in which they appear. BERT incorporates a technique called “attention mechanism”, which allows it to assign different weights to different words in the sentence, emphasizing the most relevant information. This attention mechanism helps capture long range dependencies and understand the relationships between different parts of the text. After mapping each token to its word embedding, BERT applies several layers of self-attention and feed forward neural networks to refine the representations further. These layers enable BERT to capture complex patterns and semantic relationships within the text. The final output from BERT is a length of fixed length vectors Qafeat that represent the contextualized embeddings of the input text Qaprep.

3.3 Image Features

FPN provides an effective solution for handling scale variation in images. The combination of the top-down pathway and lateral connections in FPN allows it to generate a set of feature maps, each corresponding to a specific scale. These feature maps contain multi-scale representations of image, where fine-grained details coexist with high-level semantic information. By integrating features from different scales and preserving both local and global



Article Title: Development of Counting Based Visual Question Answering System by Using Transformer and Pyramid Networks with Hybrid Deep Learning Model

information, FPN enables robust and accurate visual recognition tasks. FPN feature maps are often used in conjunction with Region Proposal Networks (RPNs) to detect objects at different scales and locations within an image.

The FPN performs training of the input images I_{input} with their corresponding targets and produces I_{feat} as the deep features from the pooling layer. These I_{feat} obtained from the pooling layer contain rich and hierarchical representation of the input images. They capture both low level details and high level semantic information from the images.

3.4 Feature Integration

The counting based VQA model proposed in this paper combines text and visual features to accurately answer counting questions. BERT method is used to extract textual features Q_{feat} , while the FPN network used to pull out the image features I_{feat} . These features are concatenated and represented as $Fa^*e_{conc} = \{Q_{feat}, I_{feat}\}$, where e ranges from 1 to E , with E representing the average count of the concatenated features.

3.5 Hybrid Model

This model combines a 1DCNN and GRU which utilizes an averaging mechanism to generate accurate answers. Fa^*e_{conc} serves as the input for both the 1DCNN and GRU networks. The combined features are then passed through the fully connected layers, which further process and transform the information. By concatenating the features, the model can effectively capture the relationships between the visual and textual modalities. The 1DCNN component processes the concatenated feature vector using 1D convolutional operations. This enables the model to capture local patterns and important information within the combined features. The output of the 1DCNN represents the learned visual-textual features. Simultaneously, the GRU component receives the concatenated feature vector and processes it sequentially to capture the temporal dependencies and semantic context within the question and visual information.

Following the processing of both the 1DCNN and GRU, an averaging mechanism is applied. The outputs from the 1DCNN and GRU networks are averaged together, resulting in a combined representation that incorporates both visual and textual information. This approach allows for effective integration of both visual and textual information, leading to accurate answer prediction for counting tasks in VQA. The Proposed model is given in Figure 2



Article Title: Development of Counting Based Visual Question Answering System by Using Transformer and Pyramid Networks with Hybrid Deep Learning Model

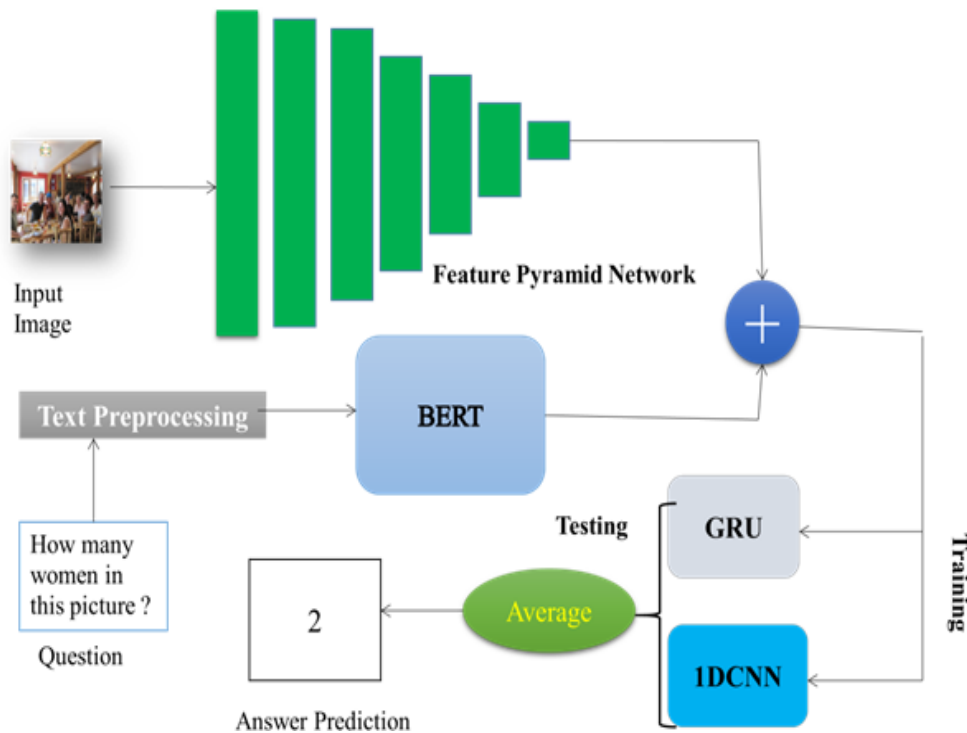


Figure 2: *Proposed model Architecture*

3 Results and Discussion

4.1 Simulation Setup

The designed counting-based VQA method Hybrid 1DCNN-GRU was implemented in Python and evaluated through experimental analysis. Various measures were employed for comparing the proposed counting-based VQA model and conventional models. The developed Proposed Hybrid 1DCNN-GRU is compared with the machine learning classifiers like LSTM, Deep CNN and RNN-CNN-MLP.

4.2 Performance metrics

Here, j is the number of fitted points, av is the actual value, fv is the forecasted value, and the value of computation is added for each fitted point is denoted by i ”.

4.2.1 MSE

MSE is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset

$$MSE = 1 / j * \sum \text{for } i \text{ to } j (av - fv)^2$$

4.2.2 RMSEs

RMSE is calculated as the square root of the mean or average of the squared differences between predicted and expected target values in a dataset.



Article Title: Development of Counting Based Visual Question Answering System by Using Transformer and Pyramid Networks with Hybrid Deep Learning Model

$$RMSE = \sqrt{1/j * \sum_{i=1}^j (av - fv)^2}$$

4.2.3 MAE

MAE score is calculated as the average of the absolute error values. Absolute or abs() is a mathematical function that simply makes a number positive. Therefore, the difference between an expected and predicted value may be positive or negative and is forced to be positive when calculating the MAE.

$$MAE = 1/j * \sum_{i=1}^j \text{abs}(av - fv)$$

4.3 Experimental Results

The resultant images along with the simple and complex questions with obtained answers from the suggested model are given in Figure 3.

Images	Questions	Actual Answers	Predicted Outcome
	How many giraffes are in the picture?	9	9
	How many bananas are on the plate?	13	13



Article Title: Development of Counting Based Visual Question Answering System by Using Transformer and Pyramid Networks with Hybrid Deep Learning Model

	How many vehicles do you see?	2	2
	How many animals are the same color?	10	10
	How many windows do you see?	7	7

Figure 3: *Generated Answers from the proposed counting-based VQA model*



Article Title: Development of Counting Based Visual Question Answering System by Using Transformer and Pyramid Networks with Hybrid Deep Learning Model

4.4 Model Evaluation

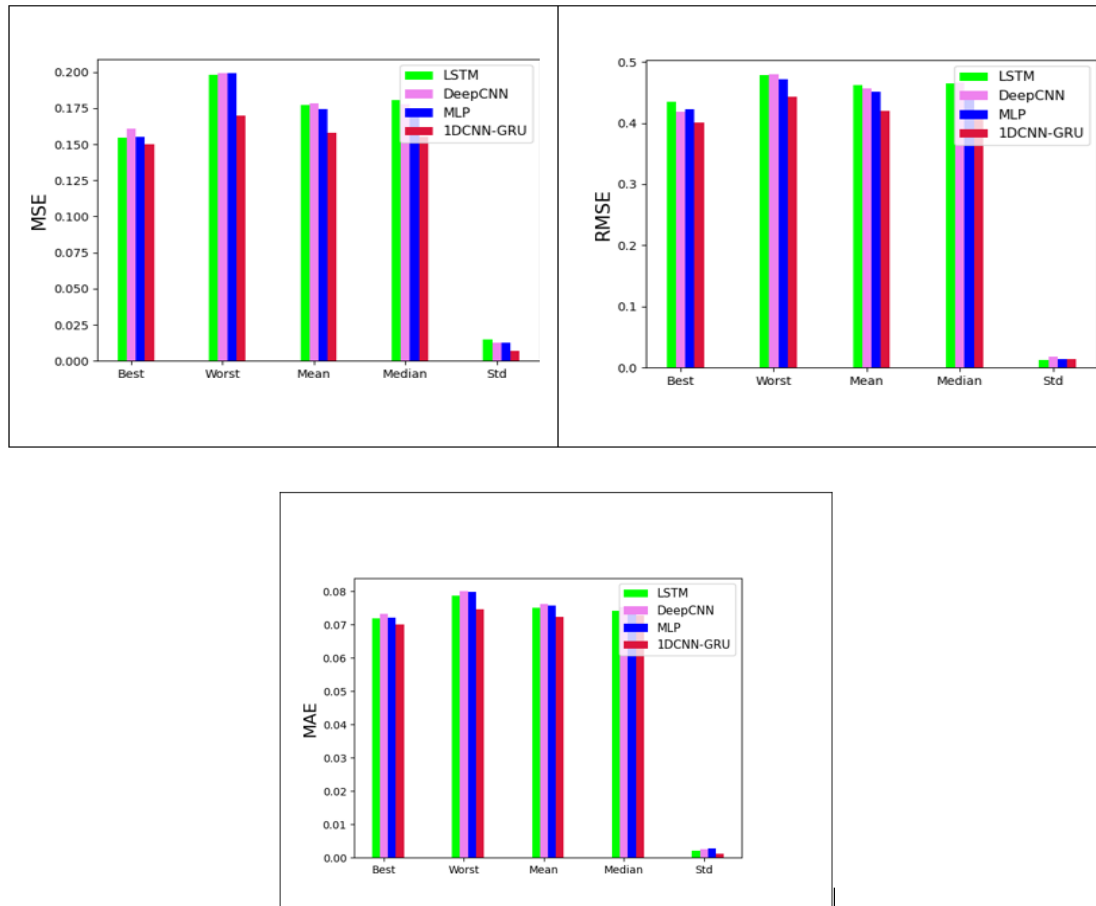


Figure 4: Statistical evaluation on designed method based on MSE, RMSE, MAE

The efficiency of the designed counting-based VQA method is evaluated with diverse deep learning methods under various statistical metrics like “best, worst, mean, median, and standard deviation” which is given in Figure 4

Table 1: Overall evaluation in all measures

MEASURES	LSTM	Deep CNN	MLP	1DCNN-GRU
MSE	0.1771	0.1785	0.17414	0.15775
RMSE	0.4617	0.4567	0.45091	0.4194
MAE	0.0750	0.07610	0.0758	0.0723

From the Table 1, we conclude that all the error measures are minimized in our proposed model



Article Title: Development of Counting Based Visual Question Answering System by Using Transformer and Pyramid Networks with Hybrid Deep Learning Model

1DCNN-GRU. Evaluation of the designed counting-based VQA model is tested based on different statistical measures with diverse. The efficiency of the designed approach starts with a lower error rate than the other classifiers and decreased severely to lower standard deviation in the MSE, RMSE and MAE analysis.

4 Conclusion

This work was designed for a counting-aided VQA method using hybrid deep learning for achieving better prediction results in the proposed model. Among all the question types, the most challenging question type is said to be counting, such as “How many?” Still, VQA models consist of certain difficulties in counting the objects that are present in the natural images. The basic technique in the VQA involved either classifying answers according to a fixed-length description of both the question and image or estimating summing fractional counts from every image section. Soft attention in these methods is utilized to find these primary issues. To circumvent this problem, the major scope of this paper is to implement a new visual question answering system based on a counting scenario.

References

1. Zhenghang Yuan; Lichao Mou; Qi Wang; Xiao Xiang Zhu, Year: 2022, “From Easy to Hard: Learning Language-Guided Curriculum for Visual Question Answering on Remote Sensing Data”, IEEE Transactions On Geoscience and Remote Sensing, Vol: 60, pp. 1 – 11.
2. Sylvain Lobry; Diego Marcos; Jesse Murray; Devis Tuia, Year: 2020, “RSVQA: Visual Question Answering for Remote Sensing Data”, IEEE Transactions On Geoscience and Remote Sensing, Vol: 58, no: 12, pp. 8555 – 8566.
3. Abhirama Subramanyam Penamakuri; Manish Gupta; Mithun Das Gupta; Anand Mishra, Year: 2023, “Answer Mining from a Pool of Images: Towards Retrieval-Based Visual Question Answering”, Computer Vision and Pattern Recognition.
4. Ming Yan; Haiyang Xu; Chenliang Li; Junfeng Tian; Bin Bi; Wei Wang; Weihua Chen; Xianzhe Xu; Fan Wang; Zheng Cao; Zhicheng Zhang; Qiyu Zhang; Ji Zhang; Songfang Huang; Fei Huang; Luo Si; Rong Jin, Year: 2021, “Achieving Human Parity On Visual Question Answering”, ACM Transactions on Information Systems, Vol: 41, no: 79 , pp. 1 – 40.
5. Ankan Bansal; Yuting Zhang; Rama Chellappa, Year: 2020, “Visual Question Answering on Image Sets”, Computer Vision – ECCV, Vol: 12366, pp. 51 – 67.
6. Sudeep D. Thepade; Jovian A. Jaison, Year: 2022, “Fusion of vectored text descriptors with auto extracted deep CNN features for improved image classification”, Intelligent Systems with



Article Title: Development of Counting Based Visual Question Answering System by Using Transformer and Pyramid Networks with Hybrid Deep Learning Model

Applications, Vol: 16, pp. 200108.

7. Zhi Lei; Guixian Zhang; Lijuan Wu; Kui Zhang; Rongjiao Liang, Year: 2022, “A Multi-level Mesh Mutual Attention Model for Visual Question Answering”, Data Science and Engineering, Vol: 7 , pp. 339 – 353.

8. Argho Sarkar; Tashnim Chowdhury; Robin Roberson Murphy; Aryya Gangopadhyay; Maryam Rahnemoonfar, Year: 2023, “SAM-VQA: Supervised Attention-Based Visual Question Answering Model for Post-Disaster Damage Assessment on Remote Sensing Imagery”, IEEE Transactions on Geoscience and Remote Sensing, Vol: 61.

9. Doaa B. Ebaid; Magda M. Madbouly; Adel A. El-Zoghabi, Year: 2023, “Bi-directional Image–Text Matching Deep Learning-Based Approaches: Concepts, Methodologies, Benchmarks and Challenges”, International Journal of Computational Intelligence Systems, Vol: 16, no: 1, pp.81.

10. Remi Cadene; Hedi Ben-younes; Matthieu Cord; Nicolas Thome, Year: 2019, “MUREL: Multimodal Relational Reasoning for Visual Question Answering”, In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, pp. 1989 – 1998.

11. Yan Zhang; Jonathon Hare; Adam Prugel-Bennett, Year: 2018, “Learning to count objects in Natural Images”, Artificial Intelligence.

12. Ander Salaberria; Gorka Azkune; Oier Lopez de Lacalle; Aitor Soroa; Eneko Agirre, Year: 2023, “Image captioning for effective use of language models in knowledge-based visual question answering”, Expert Systems with Applications, Vol: 212, pp.118669.

13. Sho Higa; Koji Yamada; Shihoko Kamisato, Year: 2023, “Intelligent Eye-Controlled Electric Wheelchair Based on Estimating Visual Intentions Using One-Dimensional Convolutional Neural Network and Long Short-Term Memory”, Sensors, Vol: 23, no: 8, pp. 4028.

14. Daniel Gordon; Aniruddha Kembhavi; Mohammad Rastegari; Joseph Redmon; Dieter Fox; Ali Farhadi., Year: 2018, “IQA: Visual Question Answering in Interactive Environments”, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4089 – 4098.

15. Zijing Yang; Cheng Wang, Year: 2023, “Short-term traffic flow prediction based on AST-



Article Title: Development of Counting Based Visual Question Answering System by Using Transformer and Pyramid Networks with Hybrid Deep Learning Model

MTL-CNN-GRU”, IET Intelligent Transport Systems.

16. Siyu Lu; Yueming Ding; Mingzhe Liu; Zhengtong Yin; Lirong Yin; Wenfeng Zheng, Year: 2023, “Multiscale Feature Extraction and Fusion of Image and Text in VQA”, International Journal of Computational Intelligence Systems, Vol: 16, no: 1, pp. 54.

17. Siyu Lu; Yueming Ding; Zhengtong Yin; Mingzhe Liu; Xuan Liu; Wenfeng Zhen ; Lirong Yin, Year: 2023, “Improved Blending Attention Mechanism in Visual Question Answering”, Computer Systems Science and Engineering, Vol: 47, no: 1.

18. Yonglin Yu; Haifeng Li; Hanrong Shi; Lin Li; Jun Xiao, Year: 2023, “Question-guided feature pyramid network for medical visual question answering”, Expert Systems With Applications, Vol: 214, pp. 119148.

19. Riccardo La Grassa; Ignazio Gallo; Nicola Landro, Year: 2020, “Learn class hierarchy using convolutional neural networks”, Computer Vision and Pattern Recognition, Vol: 51, pp: 6622 – 6632.

20. Mingqin Chen; Shan Chen; Yilei Wang; liYingjie Wu, Year: 2019, “Counting Attention Based on Classification Confidence for Visual Question Answering”, IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking, pp. 1173 – 1179.

21. Sanjay Subramanian; Medhini Narasimhan; Kushal Khangaonkar; Kevin Yan; Arsha Nagrani; Cordelia Schmid; Andy Zeng; Trevor Darrell; Dan Klein, Year: 2023, “Modular Visual Question Answering via Code Generation”, Computation and Language.

22. Abdulganiyu Abdu Yusuf; Feng Chong; Mao Xianling, Year: 2022, “An analysis of graph convolutional networks and recent datasets for visual question answering”, Artificial Intelligence Review, Vol: 55, pp: 6277 – 6300.

23. Sylvain Lobry; Diego Marcos; Benjamin Kellenberger; Devis Tuia, Year: 2020, “Better Generic Objects Counting When Asking Questions to Images: A Multitask Approach for Remote Sensing Visual Question Answering”, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol: 2, pp: 1021 – 1027.

24. Jae Won Cho; Dong-Jin Kim; Hyeonggon Ryu; In So Kweon, Year: 2023, “Generative



Article Title: Development of Counting Based Visual Question Answering System by Using Transformer and Pyramid Networks with Hybrid Deep Learning Model

Bias for Robust Visual Question Answering”, Computer Vision and Pattern Recognition.

25. Kenneth Marino; Mohammad Rastegari; Ali Farhadi; Roozbeh Mottaghi, Year: 2019, “OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge”, Computer Vision and Pattern Recognition.

26. Weidong Tian; Bin He; Nanxun Wang; Zhongqiu Zhao, Year: 2019, “Multi-Channel Co-Attention Network for Visual Question Answering”, 2020 International Joint Conference on Neural Networks.

27. S. Sarath; J. Amudha, Year: 2020, “Visual question answering models Evaluation”, International Conference for Emerging Technology.

28. Siyu Lu; Mingzhe Liu; Lirong Yin; Zhengtong Yin; Xuan Liu; Wenfeng Zheng, Year: 2023, “The multi-modal fusion in visual question answering: a review of attention mechanisms”, PeerJ Computer Science.

29. Giacomo Aldegheri; Surya Gayet; Marius V. Peelen, Year: 2023, “Scene context automatically drives predictions of object transformations”, Cognition.

30. Seung-Min Choi; Seung-Ik Lee; Jae-Yeong Lee; In So Kweon, Year: 2023, “Semantic-guided de-attention with sharpened triplet marginal loss for visual place recognition”, Pattern Recognition.

31. Caterina Trentin; Heleen A. Slagter; Christian N. L. Olivers, Year: 2022, “Visual working memory representations bias attention more when they are the target of an action plan”, Cognition.

32. Theresa Elise Wege; Bert De Smedt; Camilla Gilmore; Matthew Inglis, Year: 2022, “Counting many as one: Young children can understand sets as units except when counting”, Journal of Experimental Child Psychology.

33. Yan Gong; Georgina Cosma, Year: 2022, “Improving visual-semantic embeddings by learning semantically-enhanced hard negatives for cross-modal information retrieval”, Pattern Recognition.