IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 7, July 2024

DOI: 10.17148/IJARCCE.2024.13708

Visual Question and Answering (VQA): ViT/SwinT and BERT/RoBERTA

Adarsh Pujari[1], Digambar Dhanagar[2], Milan Srinivas[3], Aryaman Shukla[4], Rishi Singh[5]

B.Tech, Department of AI & Data Science, Sharad Institute of Technology College of Engineering, India [1]

Student, MSc (Computer Science), Scaler Neovarsity, Bangalore, India [2]

Student, BE, Department of CSE, VKIT, Bangalore, India [3]

Grade 10, High School [4]

BE, Department of Mechanical Engineering, PES (South Campus), India [5]

Abstract: In the study of artificial intelligence, Visual Question Answering is becoming a more important subject since it sits at the critical nexus of Computer Vision (CV) and Natural Language Processing (NLP). In the fields of CV and NLP, VQA has emerged as a major research area due to its cognitive capability. The semantic information needed for image captioning and video summarization is already present in still photos or video dynamics; it just needs to be extracted and articulated in a way that makes sense to humans. On the other hand, VQA doubles the effort linked to artificial intelligence by requiring semantic information from the same medium to be compared with the semantics implied by a query expressed in natural language. Transformers model is applied to the CV field and combines the transformers based NLP algorithm to construct a VQA system, based on a large number of actual scene photographs [1-3] on the KAGGLE platform. The results of the experiment validate the usefulness of their model by demonstrating that it can provide accurate answers in a simple and ordered setting and that there is a definite discrepancy between the generated results and the real answers in a chaotic scenario. We have considered the Issues or challenges based on VQA research as per the current scenario.

Keywords: Visual Question Answering (VQA), Computer Vision (CV), Natural Language Processing (NLP), Long Short-Term Memory (LSTM), MDETR, Issues or challenges based on VQA and VQA in Ontology.

I. INTRODUCTION

A. DEFINITION OF VQA

The discipline of Visual Question Answering, or VQA for short, is one of the most interesting areas of artificial intelligence. What does it involve, then? VQA is essentially concerned with a computer system's capacity to respond to inquiries regarding images. Consider posing a question about what's in a picture to a computer that you have seen it. The "Answer Image" or written response based on the system's comprehension would be provided after it analyzed the "Question Image" with the aid of AI Image Generator. With a twist on design, it resembles the conventional "Question Answering" systems. Since VQA works with images rather than just text, it is a special kind of natural language processing combined with image recognition. [2]

B. SIGNIFICANCE OF VQA

Currently, one of the most intriguing collaborative applications of Artificial Intelligence (AI) to Computer Vision (CV) and Natural Language Processing (NLP) is Visual Question Answering (VQA). The ultimate goal is to develop systems that can respond to all kinds of queries pertaining to any picture and stated in natural language. In order to solve this multidisciplinary problem, natural language processing (NLP) must first comprehend the query and then produce an answer based on the CV's results. An extended problem in NLP is text-based question and answer. With VQA, the distinction is that an image's content is taken into consideration by both search and reasoning. Automated VQA systems can also be a dependable resource for patients who need to verify medical information, follow up on past examinations, or seek professional assistance. Online Med-VQA systems can lower costs and remove physical barriers to healthcare, opening up access to a wider population. A comprehensive use of Emergency healthcare situations are where medical VQA is most useful, especially when non-experts must act quickly and efficiently. For example, if someone gets bitten by a snake, they can take a picture of the wound to confirm if the snake was poisonous. The individual may also inquire about the procedures needed to treat a poisonous snakebite. [3-6].

DOI: 10.17148/IJARCCE.2024.13708

II. Problem Statement

To build a high accuracy based Visual Question and Answering (VQA) system based on our study of research papers. In our selected study, a relatively high accuracy rate has been attained by a VQA system that is built using complicated questions and real scene photos, which explains the Attention-based Transformer approach to the Computer Vision field. In comparison to the Convolutional Neural Networks (CNN) model, this approach exhibits greater similarity between its shallow and deep network representations, and its low-level network can yield global feature information. Further evidence of the model's effectiveness in the VQA task comes from the experimental findings. To measure the effectiveness of our models, we employ the Wup similarity metric. RoBERTa-VIT [7] model has the best performance, according to the results published in the study. To summarize the findings of another study focused on training several neural networks with various topologies, including CNN and fully connected networks, CNN + Long Short Term Memory (LSTM), MobileNetv2 + LSTM, and Modulated Detection for End-to-End Multi-Modal Understanding MDETR. MDETR handled the assignment the best. When it came to counting things in a photo, [8-10] it performed poorly, but on other kinds of tests, it performed well. To address the shortcomings of the assignment with suggestions. [9-12]

III. Methodology

VQA is a challenge that combines the NLP and CV domains. The main processing notion of the VQA task is text feature extraction, image feature extraction, feature fusion, and answering based on the fused features and multi-class classifier. Figure 1 illustrates how the VQA system structure developed using CNN for cognitive answers to vision questions asked. Based on a study, a model was developed (Haiyang Tang et al., 2022) which used CNN for the VQA process.

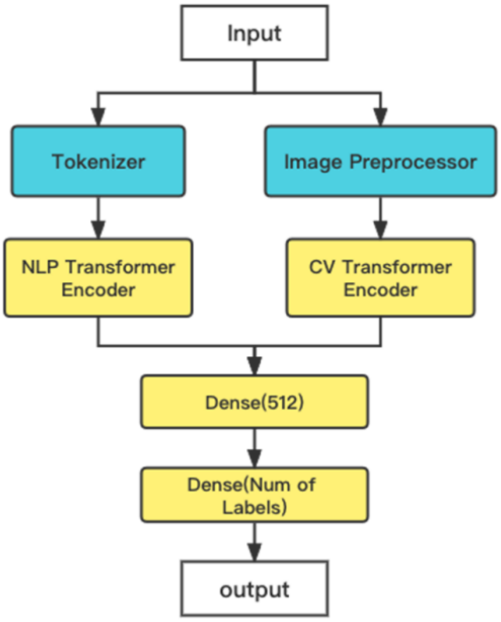


Figure 1 illustrates how the VQA system structure developed using CNN for cognitive answers to vision questions.

A. Data collection

Using PYTHON and its extensive extension library to build models, extract characteristics from images and text, and get outcomes. The KAGGLE platform provides the dataset. Kaggle is a platform for data science competitions where participants compete to create the best models for solving specific problems or analyzing certain data sets. The platform is also used for learning, collaboration, job opportunities, community building, and research in the data science and machine learning fields. In this research, the VQA dataset comprises about 1500 real-world photos along with over 12,000 text questions [13] and image-corresponding answers. Most of the real scene image information is disorganized. The data set mostly consists of disorganized bookcases, untidy bedrooms, and shopping centers filled with different goods. [9][14-16]

B. Model Selection

The models ViT/SwinT and BERT/RoBERTA are chosen to extract information from images and text, respectively, then combine the data to complete VQA tasks. The outputs are concatenated and passed through a fully-connected network with an output having the same dimensions as the answer-space after the tokenized question has been passed through the NLP Transformer Encoder and the picture features have been passed through the CV Transformer Encoder. [7][16-19]

C. Encoder

The Transformer model cuts down the distance between long-distance dependent features, resulting in more efficient feature usage and improved word-context learning. It does this by substituting the CNN network, which is frequently used in NLP applications, with the Self-Attention structure. Figure 2 shows the encoder structure of Transformer. [11] The language model then uses the transformer to create the two-way NLP model, BERT, which has a deeper understanding of the context than the one-way model. Additionally, the Masked-LM and Next Sentence Prediction pre-training tasks are the major methods used by the BERT model to enable the text semantic representation output to characterize the language. With the introduction of ViT, the boundaries between the fields of NLP and CV were broken, the conventional CNN model in CV was abandoned, and Transformer was applied to tasks such as image recognition and classification. Specifically, ViT performs feature extraction by Multi-head Self-Attention by dividing imported images into patches at specific area sizes and combining the divided patches into sequences.

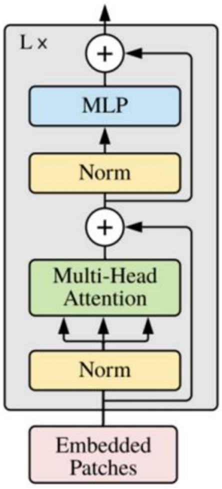
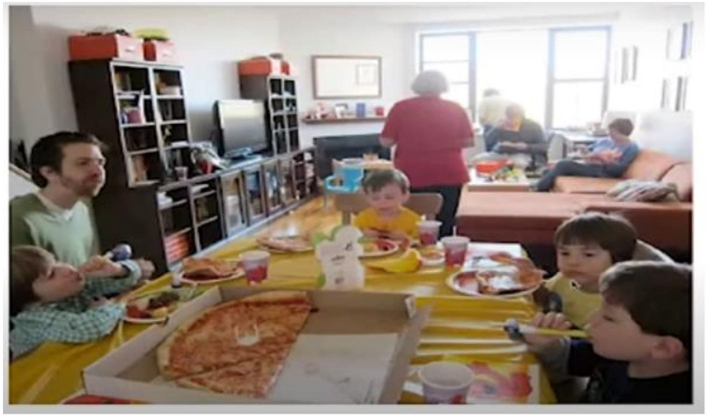


Figure 2 shows the encoder structure of the Transformer.

D. Checking the performance of the model

Wup similarity metric is used to quantify the difference between the answers and the findings in order to thoroughly assess the performance of various models in the VQA work. This metric, which can be used to indicate the similarity between two words, was proposed by Wup and is based on the WordNet dictionary. The closer the meanings are between two words or sentences, the higher the value. Figure 3 shows the VQA input image and the questions asked with answers provided by the developed model. Because the majority of the answers to the questions consists of a single word or phrase, we formed an answer-space by cutting off the first word or phrase prior to the testing and used them as labels. In this method, the VQA work is then converted into a multi-category classification problem.



Q1: How many kids are present
A: 4
Q2: How many adults are present
A: 5
Q3: Where are the kids sitting
A: on the dining table

Figure 3 shows the VQA input image and the questions asked with answers provided by the developed model. The question to confirm the number of kids in the image was asked and accurate result obtained.

DOI: 10.17148/IJARCCE.2024.13708

IV. RESULTS

The Wup similarity metric is chosen to quantify the difference between the answers and the results in order to thoroughly evaluate the performance of various models in the VQA task. (Haiyang Tang et al., 2022) This measure, which Wup suggested and is based on the WordNet database, can be used to show how similar two words are to one another. The closer the meanings are between two words or sentences, the higher the value. The RoBERTa-VIT model performs the best on the VQA task. In particular, the RoBERTa-ViT model's Wup similarity of 0.351 is 0.015, 0.007, and 0.011 higher than that of the BERT-SwinT, RoBERTa-SwinT, and BERT-ViT mixed models, in that order. Many algorithms perform comparably overall, and they all reach a respectable degree of accuracy. [20]

Evaluation of VQA by Ensembles of convolution neural networks and fully connected network, convolution neural networks and Long short-term memory, MobileNetv2 and Long short-term memory and MDETR dataset was carried out (Dmytro Koziy et al., 2021) .The data set contains images of 3D objects; each image has a number of extremely complex issues. The CLEVR data set consists of a training set of 70,000 images and 700,000 questions, a validation set of 15,000 images and 150,000 questions, A test set of 15,000 images and 150,000 questions about objects and as many answers. The MDETR model did the best job, it made a lot of mistakes when it came to counting objects in a photo, but on other types of questions, it showed phenomenal results.

V. CONCLUSIONS

A Bank proactive in business in this 21st century world has many day to day transactions. Data analytics had to be in the study (Haiyang Tang et al., 2022), they develop a VQA system using actual questions, responses, and visuals. In order to extract the features of texts and images, they integrated the BERT or RoBERTa algorithm with the Transformer model that we presented into the CV domain (VIT or SwinT) in the experiment. These outputs are then combined and sent over a network that is fully connected, resulting in an output that has the same dimensions as the input data.

The experimental results demonstrate that the model works well in straightforward and well-organized VQA tasks. The effectiveness of models is checked and verified by the possibility of a small but discernible difference between the generated results and the actual responses in a cluttered image. Wup similarity is used as a means to measure the degree of accuracy among several models. With a metric value of 0.351, the results demonstrate that the RoBERTa-ViT model performs the best in the VQA job.

The BERT-SwinT model performs comparatively poorly, which is surprising. The reason for this could be that the SwinT model is not as effective in small data set tasks as the VIT model and is more complicated. To keep raising the performance of VQA, algorithms need to be refined with more iterations contained in each epoch and expanding the dataset's size in the future.

REFERENCES

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, “Vqa: Visual question answering,” in Proceedings of the IEEE international conference on computer vision, pp. 2425–2433, 2015.

[2] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, “Visual question answering: A survey of methods and datasets,” Computer Vision and Image Understanding, vol. 163, pp. 21–40, 2017.

[3] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, “Neural module networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 39–48, 2016.

[4] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao, “Activitynetqa: A dataset for understanding complex web videos via question answering,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9127–9134, 2019.

[5] C. Unger and P. Cimiano, “Pythia: Compositional meaning construction for ontology-based question answering on the semantic web,” in International conference on application of natural language to information systems, pp. 153–160, Springer, 2011.

[6] S. Balakirsky, C. Schlenoff, S. Rama Fiorini, S. Redfield, M. Barreto, H. Nakawala, J. L. Carbonera, L. Soldatova, J. Bermejo-Alonso, F. Maikore, et al., “Towards a robot task ontology standard,” in International Manufacturing Science and Engineering Conference, vol. 50749, p. V003T04A049, American Society of Mechanical Engineers, 2017.

[7] B. Xiao, H. Wu, and Y. Wei, “Simple baselines for human pose estimation and tracking,” in Proceedings of the European conference on computer vision (ECCV), pp. 466–481, 2018.

[8] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5693–5703, 2019.

[9] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in European conference on computer vision, pp. 483–499, Springer, 2016.

[10] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “Openpose: realtime multi-person 2d pose estimation using part affinity fields,” arXiv preprint arXiv: 1812.08008, 2018.

[11] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele, “Posetrack: A benchmark for human pose estimation and tracking,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5167–5176, 2018.

DOI: 10.17148/IJARCCE.2024.13708

[12] G. Ning, J. Pei, and H. Huang, “Lighttrack: A generic framework for online top-down human pose tracking,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 1034–1035, 2020.
[13] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, “Make skeleton-based action recognition model smaller, faster and better,” in Proceedings of the ACM Multimedia Asia, pp. 1–6, 2019.
[14] H.-H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, “Skeletal movement to color map: A novel representation for 3d action recognition with inception residual networks,” in 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 3483–3487, IEEE, 2018. [15] P. Wang, Z. Li, Y. Hou, and W. Li, “Action recognition based on joint trajectory maps using convolutional neural networks,” in Proceedings of the 24th ACM international conference on Multimedia, pp. 102–106, 2016.
[16] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, “Towards understanding action recognition,” in International Conf. on Computer Vision (ICCV), pp. 3192–3199, Dec. 2013.
[17] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” arXiv preprint arXiv: 1804.02767, 2018.
[18] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in Advances in neural information processing systems, pp. 91–99, 2015.
[19] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, “Aggregated residual ´ transformations for deep neural networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492–1500, 2017.
[20] M. Tan and Q. V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” arXiv preprint arXiv: 1905.11946, 2019.