Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Enhancing VQA with SELM: A Multi-Model Approach Using SBERT

Kamala Mekala[1], Dr.Siva Rama Krishna Sarma Veerubhotla [2]
[1] *Research Scholar, kamala.mekala@gmail.com, Computer Science Engineering, Koneru Lakshmaiah Education Foundation Vaddeswaram, India*
[2] *sharmavsrk@kluniversity.in, Computer Science Engineering, Koneru Lakshmaiah Education Foundation Vaddeswaram, India*

ARTICLEINFO

Received: 29 Dec 2024
Revised: 12 Feb 2025
Accepted: 27 Feb 2025

ABSTRACT

In VQA or Visual Question Answering, a model is provided with an image and a natural language question related to it. For the model to generate appropriate answers, it must be able to understand both textual and visual input. However, there are still we have two key challenges persist in VQA.The first challenge is the inconsistency of answers and explanations provided by current approaches. The second is bridging the semantic gap that exists in between images and questions, resulting in explanations that are less accurate. Our goal is to reduce problems between image (any image) visual components and text generation alongside imbalance compensation. We propose a novel approach named System of Ensemble Learning model (SELM).The proposed approach utilizes stacked models for the extraction of text and an image features. The output of the stacked models are taken as input to the multi model fusion transformer (Similarity BERT) The SBERT model compares the predicted output with the actual ground truth results. The proposed SBERT has 95% accuracy, making it better than the state-of-the-art methods. In the future, this model may be extended to different domains like healthcare, geospatial, and satellite images etc.

Keywords: VQA ,CNN ,NLP ,SBERT ,Gemini API

1.INTRODUCTION

Visual Question Answering (VQA) is a model designed to understand both visual and textual information, which produces a response to an image and a natural language query regarding it. Despite recent advancements, two significant challenges remain: the inconsistency of answers and explanations provided by current approaches and the semantic gap between images and questions, leading to less accurate explanations. [11] Generating answers based on the contentents delived from video.[12] E-VQA datasets provides enhanced answers based on images with real world and news articles. [13] focused research on remote sensing (RS) image captioning and has concentrated on addressingrelated challenges, such as improving feature fusion techniques for better contextual comprehension,[14] has worked on enhancing memory mechanisms to preserve visual context.[15] applies summarization-driven methods for caption generation. Furthermore,. [16] explores visual question generation to amplify information and improve consistency in both questions and answers

In healthcare, VQA can support medical professionals by examining medical pictures such as X-rays and MRIs, providing insights for diagnostic purposes and improving healthcare efficiency. For instance, VQA can help radiologists identify anomalies and describe findings, thereby expediting the process for physicians. In the field of education. It is similarly helpful in the retail and E-commerce to gain understanding about products.Additionally, VQA can also contribute to security and surveillance systems by enabling real-time question answering about video streams, helping security personnel swiftly .identify potential threats or events.

By the below motivations, Figure 2 exhibits the System of Ensemble Learning model (SELM) as the article proposes for visual-based question answering . The suggested method extracts text and picture information using layered

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

models. The multi-model fusion transformer (Similarity BERT) receives the output of the stacked models as input. The actual ground truth findings are compared with the predicted results of the SBERT.

The key contributions of this paper as follows.

1. **Data Collection and Tokenization**: Datasets were sourced from torchutils, transformers were retrieved from torch vision, and processed images using PIL. The BERT model tokenizes the image inputs before forwarding them to the transformer, as LLMs require tokenized data for comprehension, ensuring seamless processing of image data.

2. **Attention Mask for Feature Extraction**: The large language model (LLM) may process the query and the image as inputs by employing an attention mask. . It extracts relevant features from the image, integrates them within the context provided by the question, and generates corresponding labels or responses, accordingly, ensuring a smooth transition from visual input to textual output.

3. Cross Entropy Loss for Efficient Accuracy Calculation: The VQA model leverages Cross Entropy Loss to evaluate the loss function, effectively measuring accuracy. By using CUDA for GPU acceleration, the model computes results significantly faster than on a CPU. The model's performance is further optimized via hyper parameters like GradScaler and the Adam optimizer, and has a learning rate of 0.001.

4. ResNet50 Performance: Our approach has been rigorously tested both theoretically and experimentally on the ResNet50 model. Notably, the proposed method achieved superior performance for the ResNet50 challenge, with experimental outcomes aligning closely with theoretical predictions.

This article's remaining sections are organized as follows. A thorough description of the SELM creation is given in Section 2. Section 3 describes the methodology. The experimental results and discussions presented in Section 4. In Section 5, the findings are finally made.

2. LITERATURE SURVEY

In order to answer questions based on images, Visual Question Answering (VQA) systems are leading the way in merging computer vision and natural language processing. Recent advancements in VQA emphasize on improving multimodal integration, enhancing robustness, and addressing biases, while challenges include handling complex questions and ensuring comprehensive image representation. [1] Introduced a deep learning and fully convolutional network which is capable of generating robust and comprehensive sentence description.[2] introduces a novel Supervised Attention-based Visual Question Answering (SAM-VQA) framework tailored for post-disaster damage assessment using remote sensing images. This framework enhances the interaction and efficiency of damage assessment by enabling users to ask natural language questions and receive relevant visual attention and answers.[3] Introduces integration of global and local feature of the image to provide comprehensive analysis of audiovisual. And AVQA also provides rules to improve the feature extraction and fusion. [4] Proposed TSE transformer with sentence embedding (TSE) to extract a double embedding representation of queries comprising keywords and medical information.

[5] Introduces the use of advanced feature extraction technique for images and questions. For image understanding, it employs the use of Faster-RCNNResNet to extract object-level region features, which provides a natural representation of images compared to traditional methods like VGG and ResNet. [6] Introduced a new task named Change Detection Visual Question Answering (CDVQA), leveraging natural language as the output in an effort to give regular users flexible access to updated information. This challenge predicts the related changes in data by using multitemporal aerial photos and a natural inquiry as inputs.[7] proposes the task of remote sensing visual question answering (VQA), which aims to make an intelligent agent answer questions about remote sensing scenes. This task combines the fields of natural language processing (NLP) and remote sensing image (RSI) processing. The proposed method that incorporates convolutional features to represent spatial information and word vectors for semantic word information. The method includes an bilinear approach and attention mechanism to improve feature alignment between words and spatial positions.

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[8] Presented anomaly-sensitive model does not considering computational cost. And also not addressing the processing efficiency using transformer based reasoning.[9] introduce DenseCapBert is an innovative VQA approach that uses dense captions to assist visual reasoning. . The authors created two comprehensive data sets using low- and high-resolution remote sensing images. These datasets consist of image-question-answer triplets, which are vital for training and testing VQA models in the context of remote sensing. [10]. Proposed model focussed only on image to image retrieval not on text based requests.[17-21] Introduces in both visual and textual modalities in order to generate more accurateVQA.[22-26] Incorporate unique techniques (such as co-attention and question-guided features) to enhance answer relevance and reasoning correctness. [27-29] introduced language-guided curriculum learning for VQA, enhancing model training on remote sensing data, while Shin-nosuke et al. [30-33]. Introduces specialized tasks like remote sensing or video VQA and limiting application to large VQA images.

These contributions collectively advance the field of remote sensing by introducing a novel VQA task, developing comprehensive data sets, applying deep learning models, and suggesting future research directions to enhance the system's capabilities.

- **Integration of Knowledge Bases**: The paper proposes future work on incorporating a knowledge base into frequently used VQA datasets.
- **Complexity in Data Set Construction**: The current approach for constructing data sets could be made more realistic and robust by including human annotators.
- **Converting Text into voice and also into various natural languages:**, In future, responses can effectively converted into voice and translating them into 100s of natural language, and vice versa.
- These research gaps highlight the areas that need further exploration to enhance the effectiveness and accuracy of VQA systems in remote sensing applications

3. METHODOLOGY

In this section, we proposed a model that can perform the given question q £ Q, which is related to an image I £ II. The objective of VQA is to train a new model p that is capable of giving a related answer a £ A. Let Ø denote the parameters for model p. The predicted answer â is attained as follows:

$$\hat{a} = \text{avg max } p\ Ø\ (a \mid I, q) \qquad (1)$$

Visual question answering is a prominent topic in the field of AI. It focuses on interpreting images to answer for specific information related to the given image and that can encapsulate all the responses in objective approach. The final results can then be compared with the predicted report of the original image. Corresponding questions can be generated from the image onwards. For this our model predicts the trained input object for the fusion of related feature extraction output. These methods typically use an object detector to extract multiple salient regions and apply the word tokenizer to process questions into distinct word tokens, enabling construction of fine-grained interactions between two modalities for robust reasoning.

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

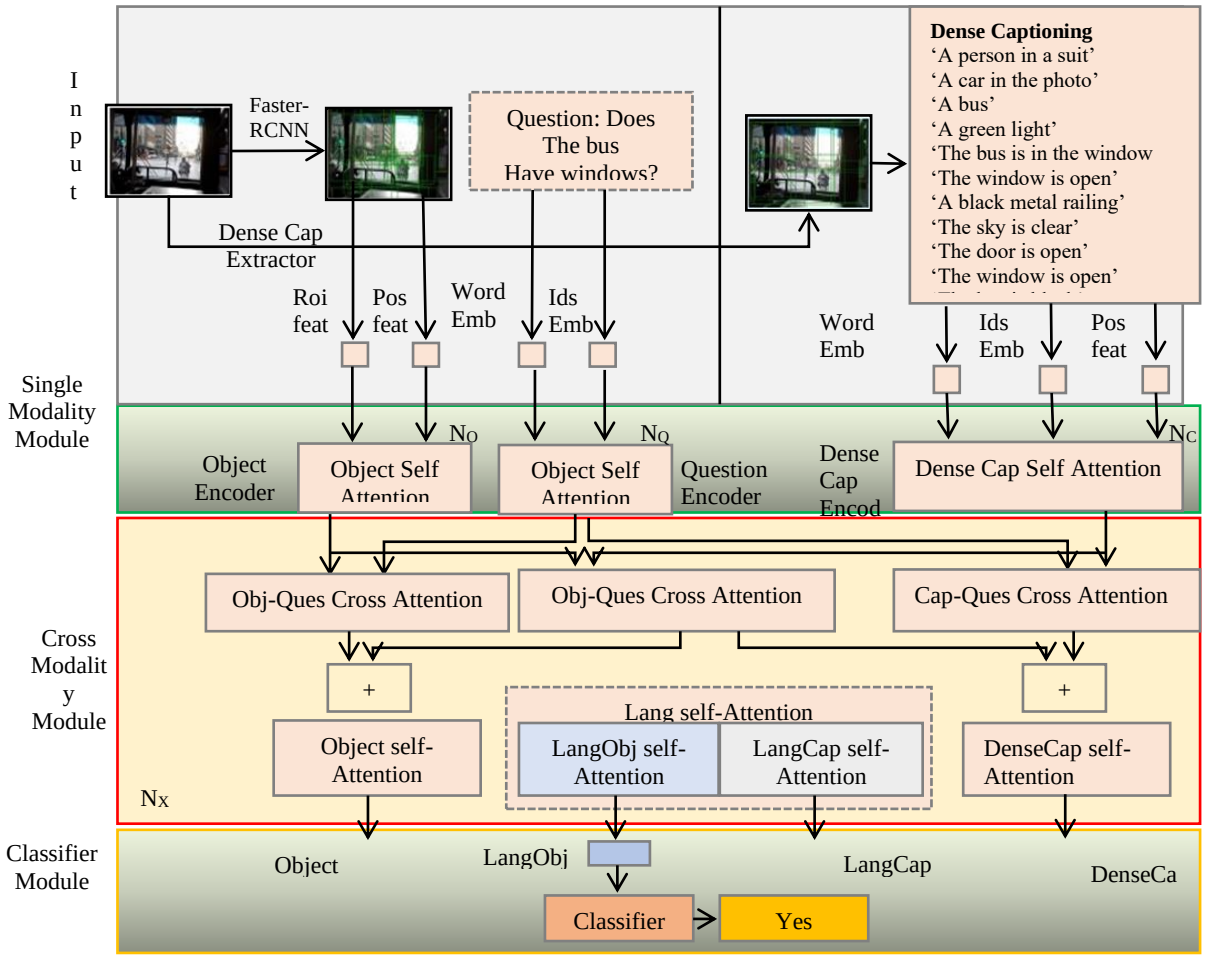


Figure 1: The architecture of the DenseCapBert model

Figure 1.The architecture of the DenseCapBert model for visual question answering.Divergingfrom conventional methods, this approach extracts dense captions from the image to enhance visual semantic information and mitigate the domain gap. Next, the Single-Modality Module to independently encodes the information from objects, questions and dense captions. We then use the proposed dense caption embedding method to encode dense captions. Then, the Cross-Modality Module formulates the fusion of information from different sources. Finally, the Classifier outputs the expected answer. NO, NQ, NC and NX represent the number of layers in the corresponding modules. Blue blocks are used to represent differences from traditional models [34].

ResNet50 uses residual connections to address the vanishing gradient problem, enabling the network to go deeper without suffering from degraded performance. The mechanism relies on residual blocks, where a skip connection (identity mapping) bypasses one or more layers, allowing the model to learn an identity function if deeper layers do not improve performance. This skip connection allows gradients to propagate back more effectively during training, thus stabilizing it for very deep networks.

$$y=F(x,\{W i\})+x \qquad (2)$$

- Here, x represents input to the residual block.
- F(x,{Wi}) indicates the series of convolutional layers applied to x.
- The output y combines the transformed input F(x,{Wi})with the original input x through an addition operation, known as the *skip connection*.

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Modeling human language on this scale is an extremely complicated and resource-intensive effort. The process of achieving the current capabilities of language models and huge language models has taken decades. Models get more complicated and effective as their size increases. Early language models could only estimate the chance of a single word; today, huge language models can predict the likelihood of phrases, paragraphs, or even complete works.

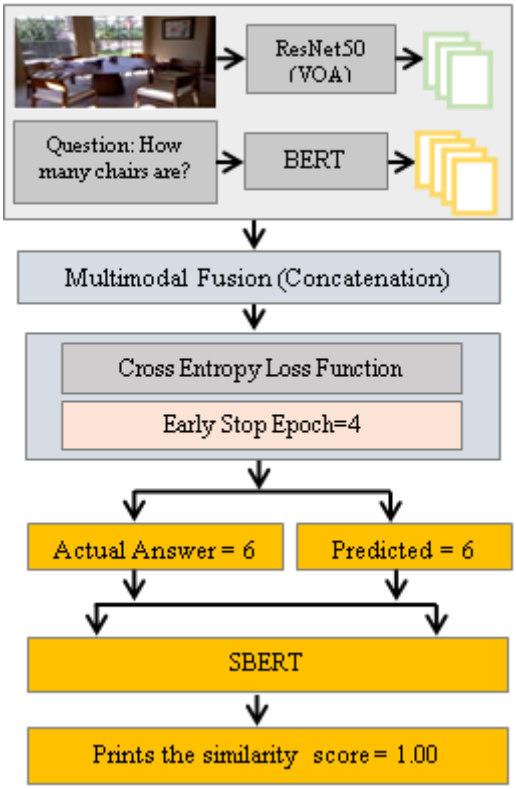


Figure 2: Architecture of the Model.

Figure 2 depict the working model of VQA which takes input as image and question and produce the result as similarity score. And is shown in detail step by step in Figure3.The qualitative comparisons is depicted in table1.

Gemini API: Gemini has access to a variety of extended dialect models. It allows you to select a LLM that most closely fits your needs, The API is capable of parsing a variety of input designs, such as chat transcripts, text portions, graphics, and code excerpts.

Convolutional Neural Networks (CNN): Convolutional neural networks (CNNs) are specialized deep learning models that are adept in processing grid-like input, such as images and videos. They consist of convolutional layers for feature extraction, pooling layers for dimensionality reduction, and fully connected layers for classification and regression applications are particularly useful in ability to recognize hierarchical patterns and attributes from basic input. Through GPU acceleration, the Gemini API may enhance CNN operations by optimizing feature extraction and increasing throughput.

Bidirectional Encoder Representations from Transformers (BERT): BERT is a sophisticated language model. This was a major breakthrough in natural language processing (NLP) with its usage of bidirectional training to anticipate words inside sentences and efficiently extract contextual meaning from both the preceding and subsequent texts. Gemini's GPU acceleration enables BERT to process queries more quickly, improving the VQA system's ability to generate responses based on analysis of complex languages.

SBERT generates embeddings: SBERT refines answer retrieval and contextual understanding by capturing semantic similarities across sentences. Gemini's GPU capabilities allow SBERT to compute embeddings rapidly, enabling effective comparison and retrieval of relevant responses. **Multimodal Fusion:** Refers to the mechanism

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

of fusing data from diverse sources to improve comprehension or study of a certain issue or occurrence. It focuses on combining information from several sources—text, photos, for example to create a more accurate and thorough depiction of the underlying facts.

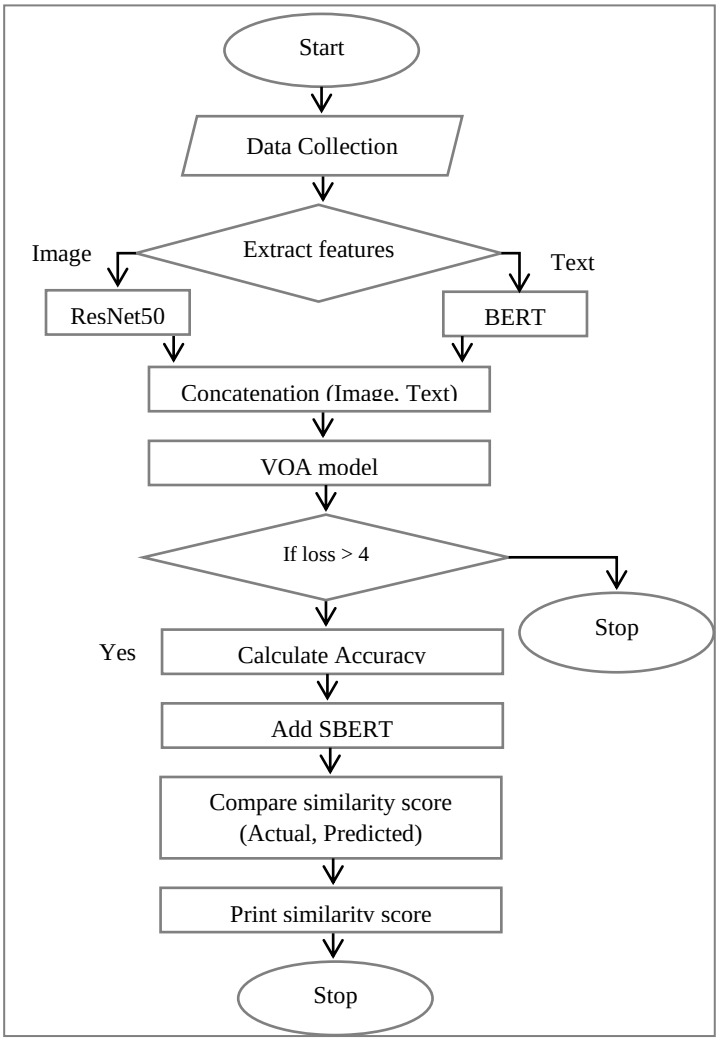


Figure3: **The flow graph for the proposed method**

Figure 3**.** Implementation of the model is shown by the Figure 3, and takes image, question as input and pre-process then it concatenates.VQA model checks if loss is less than 4 then early stop the model otherwise it calculate the accuracy and adds SBERT to compare the actual and predicted answer. Finally prints the similarity score, if both equal then score is 1 otherwise less than 1.

SBERT enhances text understanding in Visual Question Answering (VQA) by generating efficient semantic embeddings for questions, which are fused with image features for answer prediction. This improves semantic matching and question comprehension in the VQA pipeline.

Algorithm: FusionSimVQA

Start **FusionSimVQA**
Input: Dataset D1 (image features, and questions)
Output: Trained VQA model, average similarity score Sim_avg
Load dataset D and necessary libraries.
Preprocess data:

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Retain the first answer for questions with multiple answers.
Clean and normalize textual data.
Tokenize each question using the BERT tokenizer.
Extract features:
Extract image features using a pre-trained visual model.
Extract text features utilizing a pre-trained BERT model.
Concatenate image and text features into unified representations:
Split D into training and validation subsets:
Train VQA model:
Initialize model and define hyper parameters.
Train for 30 epochs, monitoring validation loss.
Apply early stopping if the validation loss does not improve after 4 epochs:
Compute similarity:
Load SBERT model.
For each validation example, calculate Sim(A, P) using cosine similarity.
Compute average similarity score Sim_avg across validation set.
Save trained model and report Sim_avg.
End **FusionSimVQA**

Table1: An overview of the most recent RSVQA methodologies' qualitative comparisons

Method	**Dataset**	**Image encoder**	**Text encoder**	**Type of method**	**Description**
Baseline(Lobry et.al.,2020)	RSVQA	Resnet-152	LSTM	VQA method	Baseline
Prompt RSVQA(Chappuris et al.,2022)	RSVQAxBEN	Resnet-152	Bert	Unified processing of models	Suitable only some specific languages
UCAGAN(Fenget al.,2023)	RSVQA	Resnet-101	LSTM	Attention method	cross-modal attention based on alternative-guided mechanism
FloodNet2.0(Sarkar et al., 2023)	FloodNet2.0			Supervised attention method	Semantic segmentation annotation required
Our Model	Resnet50	ResNet50	Roberta	VQA	Semantic matching and comprehensive answering

Table1: Shows the Qualitative analysis of analysis of existing models with proposed model including dataset, preprocessing, method used to implement and about the model .

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4. RESULTS AND DISCUSSIONS

This section outlines the evaluation metrics used by our VQA model. The study was conducted using a PyTorch framework, with a ResNet50GB GPU for training of 50 layers, ResNet50 is a deep convolution neural network design. It is well-known for its incorporation of shortcuts or skip connections to cope with the vanishing gradient issue that arises during training. This has led to the network being referred to as "Residual Network" or ResNet. These skip connection sallow the network to learn residual mappings. With its state-of-the-art performance on several benchmarks, ResNet50 has found widespread use in computer vision applications like object identification and picture categorization.

We employed Adam [34] for optimization and fine-tuning of the language model, using mini-batch sizes of 16 to a maximum of 128 epochs. Set to 2e−5, the initial learning rate decays at a rate of 1. The SBERT model employs a sequence length of 400l across all datasets. Any shorter sequence than this is padded with zero to ensure uniformity.

Our VQA algorithm generates questions, and corresponding answers based on user input, and the quality of these questions and responses are evaluated using several criteria. The initial category of metrics comprises similarity metrics, which assess the degree to which the produced questions align with the dataset's ground-truth questions.

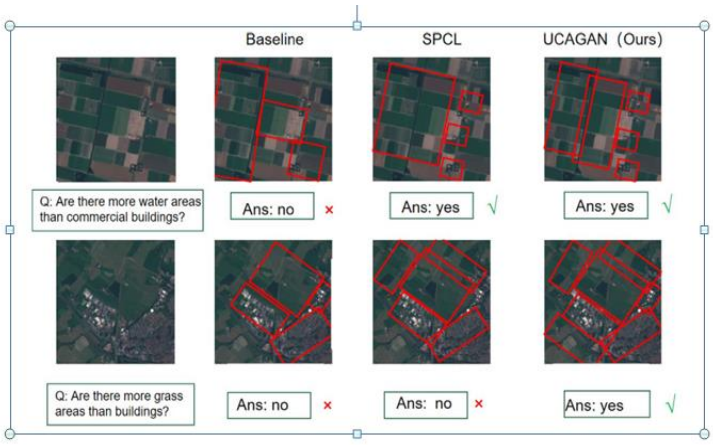


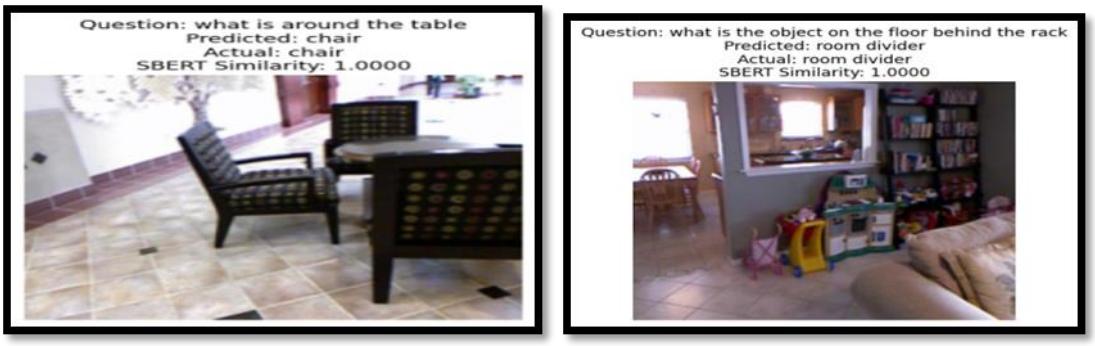
Figure 4: Comparability of VQA with existing models

Figure 4. Shows a comparison of existing models with VQA, based on the questions it is responding to, however accuracy is not aligned with its circumstance.

a b

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

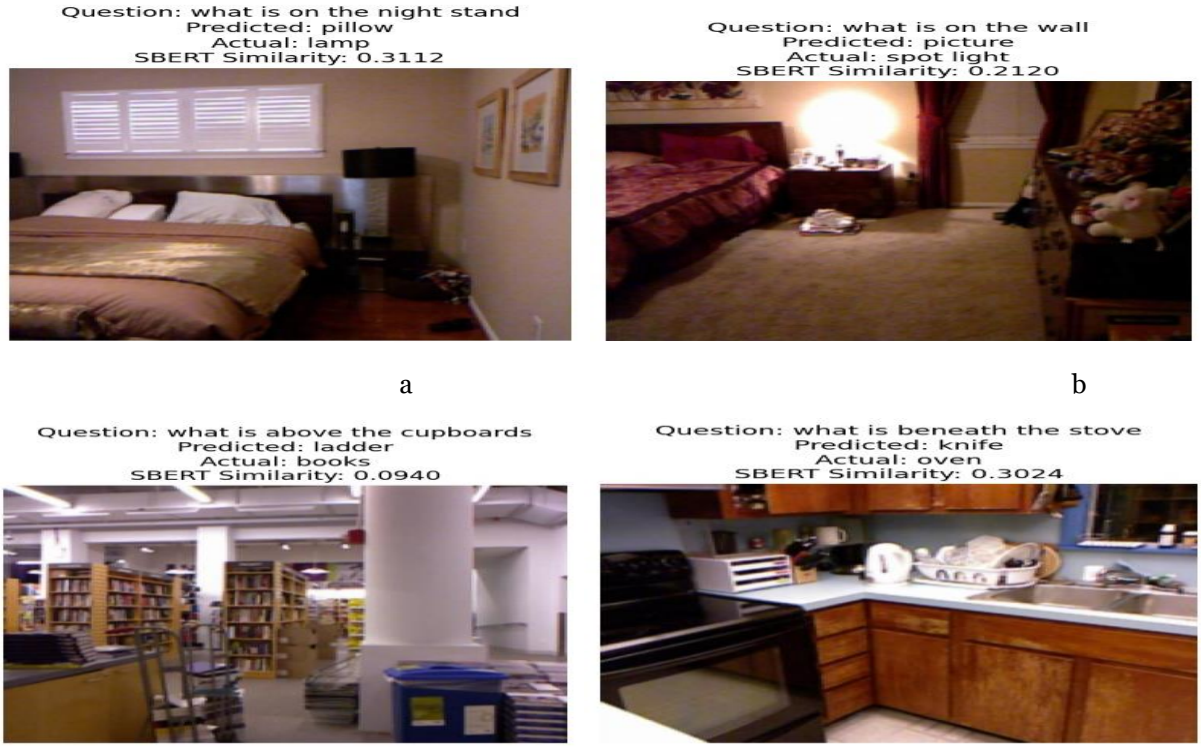
Research Article



c d

Figure 5: VQA identification using SBERT

Figure 5. Identifies the objects properly with SBERT model. The actual answer and the predicted answer both matched resulting in a similarity score of 1.



a b

c d

Figure 6: Shows low probability answer prediction

Figure 6. Shows improper results where the actual and predicted answers are not matched and similarity score is <1.Where the actual object is not matched with the predicted object.

Table 2: Comparison of SBERT with existing models results

Types	SPCL(Yuan et al.,2022)	SPCL+MLL(Yuan et al.,2022	Proposed
Object	68.91%(0.03%)	69.06%(0.13%)	81.53%(**0.51%**)
Presence	90.66%(0.08%)	91.39%(0.15%)	91.10%(**0.74%**)
Comparison	89.07%(0.27%)	89.75%(0.12%)	88.56%(**0.69%**)
Rural/Urban	85.66%(0.26%)	85.92%(0.19%)	86.83%(**0.66%**)

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Average accuracy	83.57%(0.11%)	83.97%(0.06%)	91.51%(**0.74%**)
Overall accuracy	83.09%(0.15%)	84.16%(0.05%)	95.56%(**0.85%**)

The table 2. Shows the average and overall accuracy with existing and proposed models.

Performance metrics

$$Accuracy = \frac{TP+TN}{FN+TP+FP+TN} \times 100 \qquad (3)$$

The number of objects correctly classified out of all the objects present in the test set

$$Precision = \frac{TP}{(TP+FP)} X100 \qquad (4)$$

The number of actually belongs to the positive class out of all the objects which are predicted to be of the positive class.

$$Recall = \frac{TP}{(TP+FN)} X100 \qquad (5)$$

The number of objects predicted correctly to be belonging to the positive class out of all the objects that belong to the positive class.

$$F1 - score = \frac{2(precisionXRecall)}{precision+recall} X100 \qquad (6)$$

The harmonic mean of the Precision and Recall values for the positive class.

Table 3: Comparison result with Resnet50

Method	Overall Accuracy	Precision	Recall	F1 score
SPCL+MLL(Yuan et al.,2022	83.09	75.29	86	80.28
SPCL(Yuan et al.,2022)	84.16	77.79	88.89	82.97
Proposed Model	91.39	89	96.43	92.57

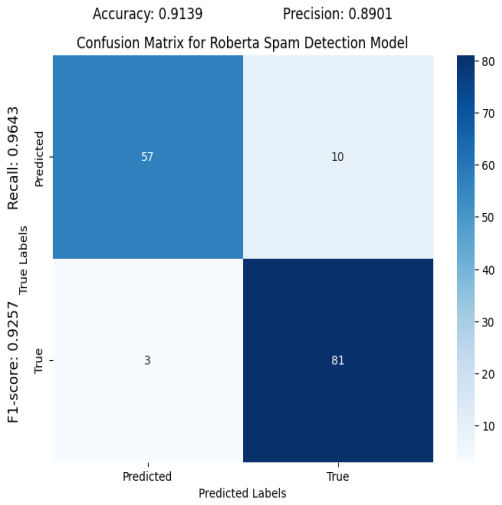


Figure 7: Confusion matrix

Figure 7. In a confusion matrix TF - 57 shows an object detection is false for actual and predicted, and TP - 81 means the actual and predicted of object in both cases is true. And FP -10 shows the actual object is false and predicted is true similarly FN -3 the actual detected object true and predicted is false

Table 3. Shows the results with overall accuracy, precision, recall and f1 score and compares with existing models with the proposed model showing the highest accuracy

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

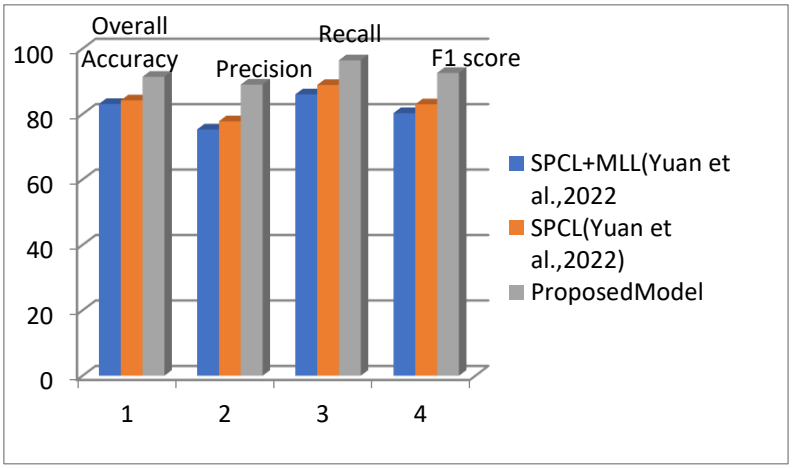


Figure 8: visualization of overall accuracy, precision, recall and f1score by column chart

Figure 8. Visualizes overall accuracy, precision, recall and f1score by using a column chart and compares two existing models with the proposed model. The green bar represents proposed data while the blue and red bars are of the existing models' information.

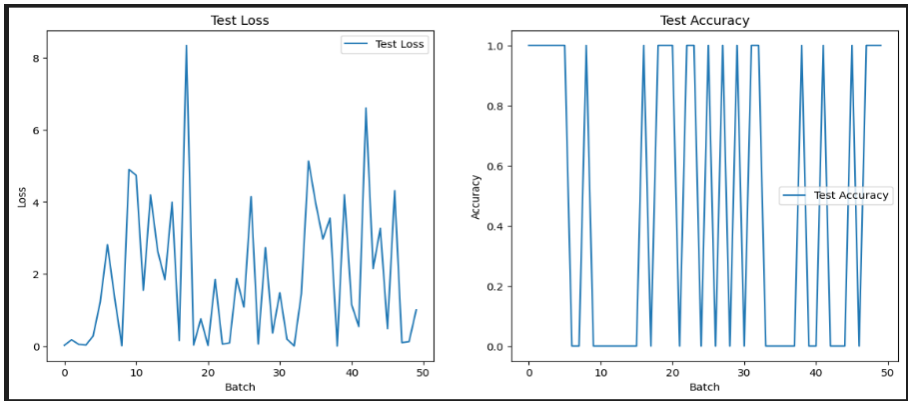


Figure 9: results of test loss, and accuracy

Figure 9. Visualizes the test loss and text accuracy is shown by plots and SBERT similarity of the Model is also represented with similarity score 1 if actual and predicted matches otherwise it is represented with <1.

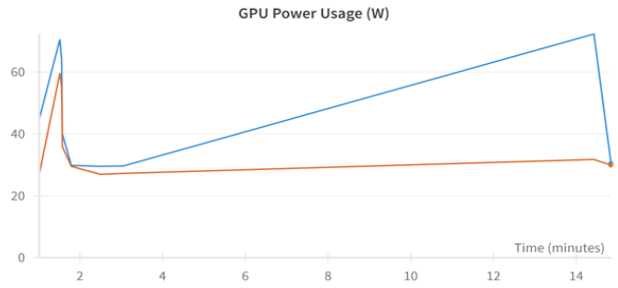


Figure 10: GPU power Usage (W)

Figure 10: GPUs is capable of handling vast volumes of data and intricate mathematical operations significantly faster than CPUs. With a high-performance GPU, tasks that would frequently take hours or even longer can be completed in just 14 minutes, demonstrating the speed of this model in comparison to other models. The blue line represents GPU time while orange represents the CPU time. The x-axis represents training time (in certain units, such as minutes, hours, or seconds) or the number of epochs and the y-axis represents the actual GPU time consumed for each epoch throughout the entire training process.

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The data represents the means of three experiments, with standard deviations shown in parentheses. In the above Figures SBERT embedded model convert the cosine value from 0 to 1.

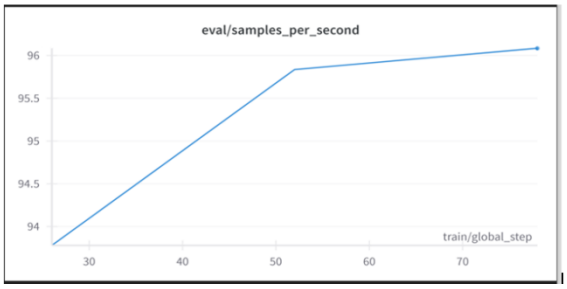


Figure 11: Evaluations samples per seconds

And the below graphs represent the status of at least 4 evolutions with epochs=30.X-axis (Epochs): This axis represent show many full iterations the model underwent during training of the whole dataset. Y-axis (Metric): the y-axis often indicates a performance metric that is being optimized or monitored, as shown in the figure 11.

Our approach yields the greatest results, enhances accuracy for all question types as well as the average and total accuracy. When compared to other authors using the same dataset implemented in the proposed SBERT algorithm. This algorithm demonstrates significantly accurate results with a accuracy of 95%, as illustrated in figure 11. This can be attributed to our method which greatly enhances access to more direct comprehension and application of natural language data, alongside reasoning and understanding regarding the content of remotely sensed pictures and the context of high-level semantics. Our method shows high performance on unbalanced samples and may be used to other datasets with similar characteristics. Future research will focus on expanding its usage with various data deployment models up to 2024 and training the model to be updated with new information. While our current approach increases the accuracy. This product may be used in live cameras to identify items or objects in the RS images. It can also be used in public areas to identify strangers or in law enforcement to identify thieves. Lastly, but just as importantly, it can be used in image analysis. This study leverages user reaction to generate questions and answers. Limitations: As pretrained model require less calculations and infrastructure, we evaluated it using a few pretrained large language models in this research.

5.CONCLUSION

In this article, the proposed model is composed using a system of ensemble learning models (SELM), which leverages stacked models for feature extraction from both image and text, combined with multi model fusion transformer (SBERT), addressing challenges like inconsistency of responses and explanations produced by existing models, as well as the semantic gap between visual and textual components. The proposed method has demonstrated its efficacy by achieving an accuracy of **95%**, significantly outperforming state-of-the-art methods. By integrating advanced ensemble learning techniques with SBERT, SELM ensures more consistent and accurate results, bridging the semantic gap and improving explanatory quality. Additionally, the model employs imbalance compensation mechanisms; effectively mitigate discrepancies between visual features and textual representations.

The results highlight the potential of SELM in not only in advancing VQA performance but also as a foundation for extending similar methodologies to other domains. Future work will focus on adapting SELM to handle domain-specific challenges, incorporating real-world noisy data, and exploring its scalability for large-scale datasets. Furthermore, research will include the generateon of answers for corresponding image and question in various languages.

REFRENCES

[1]. Shi, Zhenwei, and Zhengxia Zou. "Can a machine generate humanlike language descriptions for a remote sensing image?." IEEE Transactions on Geoscience and Remote Sensing 55.6 (2017): 3623-3634.

[2]. Sarkar, Argho, et al. "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* 61 (2023): 1-16.
[3]. Chen, Zailong, et al. "Question-aware global-local video understanding network for audio-visual question answering." *IEEE Transactions on Circuits and Systems for Video Technology* (2023).
[4]. Huang, Xiaofei, and Hongfang Gong. "A dual-attention learning network with word and sentence embedding for medical visual question answering." *IEEE Transactions on Medical Imaging* (2023).
[5].Peng, Liang, et al. "Answer again: Improving VQA with cascaded-answering model." *IEEE Transactions on Knowledge and Data Engineering* 34.4 (2020): 1644-1655.
[6].Yuan, Zhenghang, et al. "Change detection meets visual question answering." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022): 1-13.
[7].Zheng, Xiangtao, et al. "Mutual attention inception network for remote sensing visual question answering." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021): 1-14.
[8]. Cong, Fuze, et al. "Anomaly matters: An anomaly-oriented model for medical visual question answering." *IEEE Transactions on Medical Imaging* 41.11 (2022): 3385-3397.
[9]. Bi, Yandong, et al. "See and learn more: Dense caption-aware representation for visual question answering." *IEEE Transactions on Circuits and Systems for Video Technology* (2023).
[10]. Abdullah, Taghreed, et al. "TextRS: Deep bidirectional triplet network for matching text to remote sensing images." *Remote Sensing* 12.3 (2020): 405.
[11].Khurana, Khushboo, and Umesh Deshpande. "Video question-answering techniques, benchmark datasets and evaluation metrics leveraging video captioning: a comprehensive survey." *IEEE Access* 9 (2021): 43799-43823.
[12]. Yang, Zhenguo, et al. "Event-oriented visual question answering: The E-VQA dataset and benchmark." *IEEE Transactions on Knowledge and Data Engineering* 35.10 (2023): 10210-10223.
[13].Huang, Wei, Qi Wang, and Xuelong Li. "Denoising-based multiscale feature fusion for remote sensing image captioning." *IEEE Geoscience and Remote Sensing Letters* 18.3 (2020): 436-440.
[14].Fu, Kun, et al. "Boosting memory with a persistent memory mechanism for remote sensing image captioning." *Remote Sensing* 12.11 (2020): 1874.
[15].Sumbul, Gencer, Sonali Nayak, and Begüm Demir. "SD-RSIC: Summarization-driven deep remote sensing image captioning." *IEEE Transactions on Geoscience and Remote Sensing* 59.8 (2020): 6922-6934.
[16].Krishna, Ranjay, Michael Bernstein, and Li Fei-Fei. "Information maximizing visual question generation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
[17]. Bi, Yandong, et al. "Fair Attention Network for Robust Visual Question Answering." *IEEE Transactions on Circuits and Systems for Video Technology* (2024)..
[18]. Guo, Wenya, et al. "Re-attention for visual question answering." *IEEE Transactions on Image Processing* 30 (2021): 6730-6743.
[19]. Liu, Yang, Guanbin Li, and Liang Lin. "Cross-modal causal relational reasoning for event-level visual question answering." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.10 (2023): 11624-11641..
[20]. Mishra, Aakansha, Ashish Anand, and Prithwijit Guha. "Dual attention and question categorization-based visual question answering." *IEEE Transactions on Artificial Intelligence* 4.1 (2022): 81-91.
[21]. Yu, Jing, et al. "Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval." *IEEE Transactions on Multimedia* 22.12 (2020): 3196-3209.
[22].Yang, Yi, and Shawn Newsam. "Bag-of-visual-words and spatial extensions for land-use classification." *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. 2010.
[23].Cheng, Gong, Junwei Han, and Xiaoqiang Lu. "Remote sensing image scene classification: Benchmark and state of the art." *Proceedings of the IEEE* 105.10 (2017): 1865-1883.
[24]. Vu, Minh H., et al. "A question-centric model for visual question answering in medical imaging." *IEEE transactions on medical imaging* 39.9 (2020): 2856-2868.
[25]. Yang, Chao, et al. "Co-attention network with question type for visual question answering." *IEEE Access* 7 (2019): 40771-40781.

Research Article

[26].Zhang, Haonan, et al. "Learning visual question answering on controlled semantic noisy labels." *Pattern Recognition* 138 (2023): 109339.
[27].Ouyang, Ninglin, et al. "Suppressing biased samples for robust VQA." *IEEE Transactions on Multimedia* 24 (2021): 3405-3415.
[28].Liu,et al. "Inverse visual question answering: A new benchmark and VQA diagnosis tool." *IEEE transactions on pattern analysis and machine intelligence* 42.2 (2018): 460-474.
[29]. Gao, Difei, et al. "Learning to recognize visual concepts for visual question answering with structural label space." *IEEE Journal of Selected Topics in Signal Processing* 14.3 (2020): 494-505.
[30].Ishikawa, Shin-nosuke, et al. "Example-based explainable AI and its application for remote sensing image classification." *International Journal of Applied Earth Observation and Geoinformation* 118 (2023): 103215.
[31]. Qian, Tianwen, et al. "Locate before answering: Answer guided question localization for video question answering." *IEEE Transactions on Multimedia* (2023).
[32]. Mao, Aihua, et al. "Positional attention guided transformer-like architecture for visual question answering." *IEEE Transactions on Multimedia* 25 (2022): 6997-7009..
[33]. Feng, Jiangfan, et al. "Improving visual question answering for remote sensing via alternate-guided attention and combined loss." *International Journal of Applied Earth Observation and Geoinformation* 122 (2023): 103427.
[34]. Kingma, Diederik P. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).