Journal of Innovations in Computer Science and Engineering
Shahid Beheshti University

June 2025, Volume 3, Issue 1

Improving Predicted Answer Accuracy in Visual Question and Answer Systems using Attention Mechanisms and Neural Networks

Fatemeh Rezaei, **Code Orcide:**
Department of Computer Engineering, Golestan University, Gorgan, Iran, Fatemeh.rezaei1998@gmail.com,
Soheila Karbasi⌧, **Code Orcide:**
Department of Computer Engineering, Golestan University, Gorgan, Iran, s.karbasi@gu.ac.ir

Abstract
In recent years, one of the most widely studied areas in computer vision and natural language processing (NLP) is the interdisciplinary problem of Visual Question Answering (VQA), which involves the integration of computer vision and NLP. In VQA systems, an image and a text-based question about the image serve as inputs, and the system must predict the correct answer to the question. The main objective of these systems is to maximize the accuracy of correct answer predictions.

Important challenges in this field include the need for large and suitable datasets as well as powerful hardware for training the model. Key factors to improve the performance of these models include selecting the appropriate neural network for processing the inputs, selecting the appropriate dataset, and the method of combining the features extracted from the inputs. Also, using different attention mechanisms can improve the overall performance of the system. Furthermore, incorporating various attention mechanisms into the model can significantly enhance the overall performance of VQA systems. In these systems, different neural networks are employed to process inputs: convolutional neural networks (CNNs) with various architectures are used for image processing, and different types of recurrent neural networks (RNNs) are used for text processing.

In this research, the architecture of the convolutional neural network is changed and the self-attention mechanism is used in text processing and the Skipgram language model is used for embedding the input text. The performance of the proposed model is evaluated on two datasets, VQA 1.0 and VQA 2.0. The results show that the proposed model has been able to increase the overall accuracy in the VQA 1.0 dataset to 67.25% and in the VQA 2.0 dataset to 61.57%, showing significant improvement over the baseline models.

Keywords: visual question and answer system, convolutional neural network, recurrent neural network, attention mechanism.

1. Introduction

Currently, one of the most widely researched topics in artificial intelligence is VQA. VQA involves integrating both computer vision and natural language understanding (NLU). In these systems, an image and a text-based question about the image are used as inputs, and the system is responsible for automatically answering the question and predicting the correct answer. The integration of computer vision and NLU is crucial for this task [1]. The use of appropriate neural networks to process input data, along with selecting suitable datasets, are two critical factors that significantly enhance the accuracy of final predictions. Furthermore, how to integrate the features and representations obtained from inputs is a vital issue in these systems. Due to the high efficiency, importance, and attractiveness of this topic, many researchers have conducted studies to improve the accuracy of these systems, yielding positive outcomes. Various neural networks can be utilized in VQA systems. Both image processing and text processing benefit from different neural networks to ensure high accuracy of the outputs. Their efficiency in processing inputs has led to their widespread adoption. In addition, different neural networks, such as CNNs and various types of RNNs, different attention mechanisms and transformers can be used to increase output accuracy. Many studies show that these approaches significantly improve the accuracy of predicted answers in VQA models [2, 3]. Some of the challenges in this field include the need for appropriate hardware to train the models, the integration of features and representations derived from the question and image to achieve a common representation, and the selection of suitable neural networks and datasets. Today, computer vision and NLP play a crucial role across various domains, assisting human beings in diverse fields. As mentioned earlier,

Submit Date: 2025-09-22
Revise Date: 2025-11-02
Accept Date: 2025-12-07
⌧ Corresponding author

VQA requires the integration of computer vision and NLP and has numerous applications in different areas. For example, in the medical industry, VQA systems can be used to facilitate certain tasks, making processes easier for both doctors and patients.

Additionally, these systems can be highly beneficial for blind and visually impaired individuals who struggle to recognize details in images, helping them comprehend visual content and extract important information. More broadly, VQA systems are essential for tasks that require image-based information retrieval, where users can extract relevant details through natural language questions. Figure 1, illustrates the general schematic of a visual question answering model. This model takes an image and a question as inputs and predicts the appropriate answer.

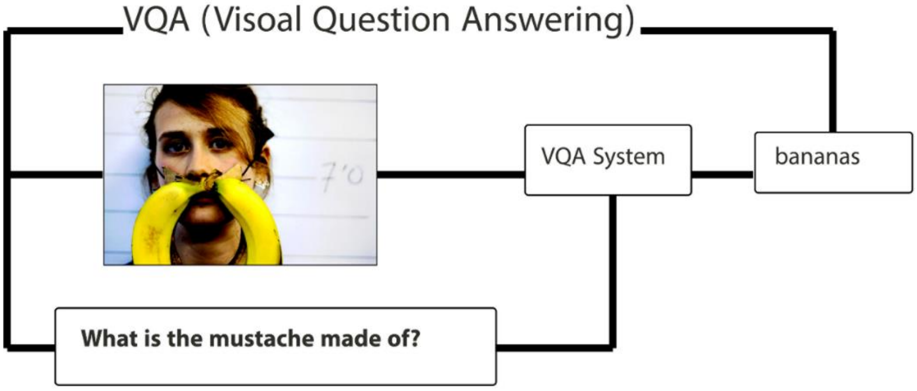
VQA (Visoal Question Answering)
VQA System
bananas
What is the mustache made of?

Figure 1: General schematic of visual question and answer system [3]

While the VQA dataset has enabled significant progress in visual question answering, ongoing researches continue to enhance the effectiveness of reasoning capabilities, linguistic diversity, and multimodal understanding. This study contributes a model based on recent advances in attention mechanisms and deep visual representation.

1.1 Challenges

In recent years, deep learning has excelled in solving various types of problems. The concept of deep learning is inspired by the functioning of the human brain. With the help of new technologies, it has achieved various successful results in artificial intelligence and machine learning. For deep learning, datasets and neural networks are two essential components [4]. Word embedding is a dense representation of words in the form of numerical vectors that can be learned by various linguistic models. Word embedding representations can reveal many hidden relationships between different words [5]. The objective function of word2vec ensures that words with the same concept and meaning have similar embeddings [6, 7]. This algorithm uses the following two methods to create word vectors:

Continuous Bag of Words (CBOW): When this method is applied to all the words in the text, the final word vectors become the target vectors [6].

Skip-gram: Algorithmically, the CBOW and Skip-gram methods are similar. The main difference is that CBOW predicts the target words from the input context, while Skip-gram works in the opposite direction, predicting context words from the target word [6].

Convolutional Neural Networks (CNNs) are among the most influential architectures in deep learning, particularly for visual recognition tasks. A standard CNN consists of convolutional layers, pooling layers, and fully connected layers, each responsible for extracting and transforming features at different levels of abstraction. While traditional CNNs operate in a purely feedforward manner, feedback CNNs introduce recurrent or top-down connections that allow the network to refine its internal representations over multiple iterations. This feedback mechanism enables the model to incorporate contextual information and adjust its predictions dynamically, which has shown to improve performance in complex tasks such as object recognition, scene understanding, and segmentation. These enhancements are particularly valuable in scenarios where initial predictions may be ambiguous or incomplete, allowing the model to "rethink" its interpretation of the input data [8, 9]. This network was inspired by experiments on the visual cortex, with its architecture simulating the pattern of connections between neurons in the brain. The features obtained by the convolutional layers are combined by the fully connected layer to form a feature vector, which is then passed to a Softmax function to predict the correct class [10]. The number of outputs of this layer matches the number of available classes [11]. Batch normalization is typically applied after a convolutional or fully connected layer and before the non-linear activation function. As discussed by Szandała [12], the placement of activation functions significantly influences the performance of deep neural networks. Batch normalization helps stabilize the distribution of inputs to these activation functions, which in turn improves their effectiveness. By reducing internal covariate shift, batch normalization enables the use of higher learning rates and mitigates sensitivity to weight initialization, ultimately accelerating training and enhancing model accuracy.

Although batch normalization is widely used to stabilize and accelerate training in deep neural networks, its direct impact on final model accuracy remains a subject of debate. As reviewed by Szandała [12], the effectiveness of activation functions is closely tied to the distribution of their inputs, which batch normalization helps regulate. However, some studies suggest that the primary benefit of batch normalization lies in enabling higher learning rates and smoother optimization, rather than consistently improving accuracy. Therefore, while BN contributes to training efficiency, its influence on accuracy may vary depending on the architecture and task.

Recurrent Neural Networks (RNNs), with their capacity to capture sequential dependencies, are widely applied in natural language processing (NLP), speech recognition, and other tasks involving ordered input data. While Li et al. [13] utilized n-gram-based models instead of RNNs, their study underscores the significance of sequence modeling in producing coherent image descriptions. While RNNs are adept at learning short-term dependencies, they struggle with long-term time series. To address this issue, the Long Short-Term Memory (LSTM) architecture was introduced by modifying the network's design [14]. LSTMs can process and classify long-term data that requires historical information and time series with delays. This network trains models using backpropagation [14].

Bidirectional Long Short-Term Memory (Bi-LSTM) is a type of RNN. Unlike unidirectional LSTMs, Bi-LSTMs process data in two directions by utilizing two hidden layers. This network maintains the temporal order of words in a text, enabling it to ignore unnecessary words using the forget gate [15]. Currently, the attention mechanism is used in both NLP and computer vision. For instance, when people read a text to understand, they do not pay attention to every word but focus on specific words to grasp the meaning. Accordingly, the attention mechanism was developed to enable neural networks to mimic human behavior [16]. By using the attention mechanism, the points and areas of an image related to a question can be identified as prominent. Further processing by neural networks on these points and areas allows for predicting the appropriate answer. Additionally, the attention mechanism can be used to better process input text in VQA systems. In fact, models using the attention mechanism to solve image question-answer problems can apply this mechanism to the image, the question, or both simultaneously [17]. In VQA systems, an image and a question in natural language are provided as inputs. The VQA system attempts to predict the appropriate answer based on these inputs, using the visual elements of the image and inferences derived from the question [18]. The main challenge in these systems is successfully combining the representations produced for the input image and question using different neural networks to predict the correct answer [19].

Approaches based on hybrid models with modular structures enable the determination of appropriate architectures to answer questions according to the incoming query. Among these hybrid models, we can mention those based on dynamic memory networks (DMNs) [19].

The QA-COCO dataset includes 123,268 images, with 72,783 used for training and 38,948 for testing. Each image is paired with a question and an answer. The questions in this dataset fall into four categories: object, color, number, and location. Six percent of the questions in the QA-COCO dataset have one-word answers, which simplifies evaluation. However, one of the limitations of this dataset is the presence of numerous grammatical errors in the questions [20].

The VQA 1.0 dataset represents the initial release of the Visual Question Answering benchmark, utilizes images sourced from the QA-COCO dataset. It comprises 204,721 images, each paired with three questions. For every question, ten human-provided answers are included, enabling robust evaluation of model performance across diverse linguistic interpretations [21].

The VQA 2.0 dataset is the second version of the VQA dataset. It also contains 204,721 images from the QA-COCO dataset, but includes 110,504 questions. There are between three and five questions per image, with ten answers available for each question [21].

2. Related works

Ma et al. [22] proposed a CNN-based framework to predict answers in Visual Question Answering (VQA) systems, demonstrating the effectiveness of convolutional features in understanding image content and aligning it with natural language queries. Unlike most models in this field, the proposed model does not use RNNs, but relies only on a set of CNNs. This model comprises three CNNs. The first CNN extracts some features from the input image using a 16-layer Net-VGG type architecture [8] to generate a representation of the image. Another CNN processes the input text where initially, the words (questions) of the input text are embedded using the Skipgram model [6]. Then, convolutional layers followed by max-pooling layers process the input text, extract its features, and generate a representation of the text. The schematic of the CNN used for text processing is shown in Figure 2.

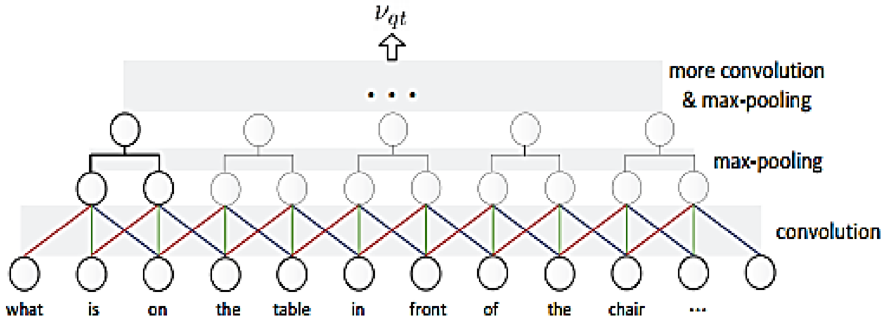

Figure 2: Convolutional neural network (CNN) schematic for input text processing [22]

Using these representations, a multi-modal CNN generated a joint embedding by combining features from both the input text and image. In this framework, ResNet [10] was employed as the visual feature extractor. The resulting fused representation was passed through a Softmax layer to predict the most probable answer. To evaluate the performance of this model, the QA-COCO dataset and DAQUAR datasets [20] were used. This model achieved 54.40% accuracy on the QA-COCO dataset and 42.76% accuracy on the DAQUAR dataset for single-word responses. The framework of the proposed model is shown in Figure 3.

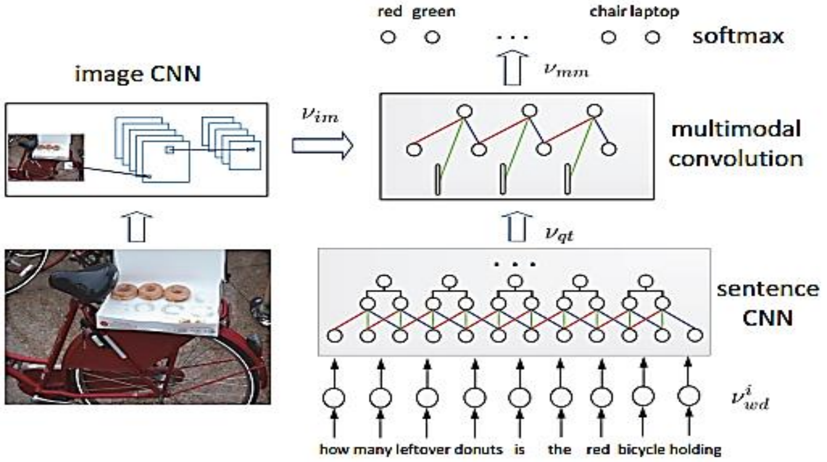

Figure 3: Architecture of the proposed model to predict the response [22]

Malinowski et al. [23] proposed a neural-based framework that combined Convolutional Neural Networks (CNNs) for visual feature extraction with Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) units, to model sequential dependencies in textual questions. This architecture enabled the system to jointly reason over image and language modalities, effectively predicting answers in Visual Question Answering (VQA) tasks. In the proposed method, a pre-trained CNN (GoogLeNet type) is used to extract features from the input image. Because the final answer may consist of multiple words, in each iteration, the last predicted word is fed into the short-term memory network through a recursive loop to predict the subsequent words. The schematic of the answer generation process in the proposed method is shown in Figure 4.

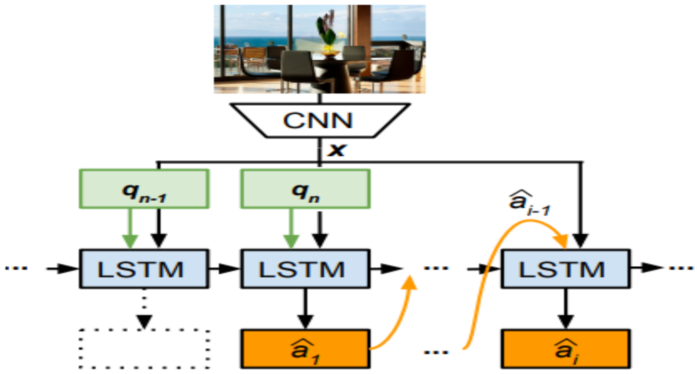

Figure 4: Prediction of the response in the proposed model [23]

A new framework based on the attention mechanism called "attention co-Hi" is proposed to predict answers in VQA systems [24]. In this method, three stages of embedding are used on the input question.

A new framework based on dynamic memory networks (DMNs) is proposed for solving the image question answering problem [25]. This model employs four modules: input module, question module, episodic memory module, and answer module. The input module processes inputs with different neural networks and converts them into vectors called "Facts". The input to this module is the image, which is divided into small local areas, each treated as equivalent to a sentence in the text input module.

A Bi-LSTM network is used to process the input text, and a Faster R-CNN neural network processes the input image, predicting the answer in the VQA system [26].

A novel approach (GloVe) is introduced to learn **word embedding** by combining **global matrix factorization** and **local context window methods [7]**. Unlike traditional models, GloVe constructs word vectors based on **word co-occurrence statistics**, capturing fine-grained semantic relationships. The authors propose a **log-bilinear regression model** that efficiently leverages statistical information by training only on **nonzero elements** in a word-word co-occurrence matrix. This method results in a **vector space with meaningful substructure**, enabling **word analogy tasks** and outperforming previous models in **similarity tasks and named entity recognition**.

A novel approach is presented to **Zero-Shot Learning (ZSL) [27]**. One of the key challenges in visual classification tasks is the difficulty of collecting training images. Traditional ZSL methods often rely on **textual descriptions or attribute-based representations** to define new categories. However, this study introduces an **alternative modality** by leveraging **visual abstraction** to learn complex concepts. This research specifically focuses on **human-related concepts and interactions**. The authors propose a method that allows users to **generate training data through abstract visualizations,** such as **clip-art representations** where characters' **body posture, facial expressions, gaze, and gender** can be manipulated to depict interactions. This approach enables **models to be trained on abstract images** and later tested on real-world photographs. The findings of this study demonstrate that **visual abstraction** can be effective in learning complex concepts and can outperform certain traditional approaches. Additionally, the paper introduces an **explicit mapping between abstract and real-world domains**, helping models bridge the gap.

A number of neural network modules and their combinations are employed in a new model to predict appropriate responses in visual question answering (VQA) systems [28]. In this model, after embedding the words of the input text, the text is passed through a recurrent neural network (RNN) based on long short-term memory (LSTM), where the question is processed, and an output is generated. Additionally, the embedded text is fed into a text parser that decomposes the text into its fundamental components and analyzes them. Based on the output of this analysis, the model selects one or a combination of modules needed to predict the response. For image processing, the model utilizes a convolutional neural network (CNN) with a 16-layer VGGNet architecture to extract visual features. These extracted features serve as input to the existing modules, which operate based on an attention mechanism. Finally, the output of a single module or a combination of modules is integrated with the text representation, and the final answer is predicted using Softmax layers. The proposed architectural structure of this study is illustrated in Figure 5.

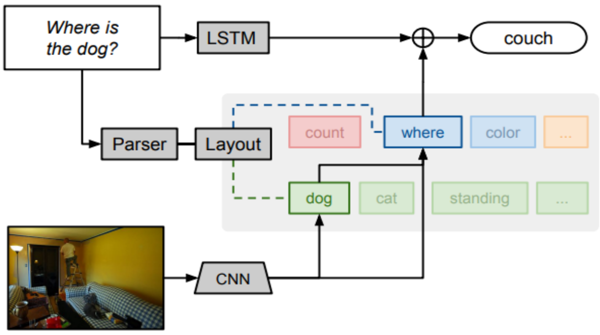

Figure 5: The proposed architectural structure [28]

In recent years, **bounding box-based features** have become the standard in vision-and-language tasks like VQA. The effectiveness of **grid features** in **VQA systems is** examined [29] which reevaluates **grid features** and demonstrates their significant potential. This research found that **grid features** not only offer **higher processing speed** but can also achieve **comparable accuracy** to bounding box-based features when properly preprocessed [29]. Through **extensive experiments**, the study confirms that these findings hold across different VQA models, various datasets, and even in related tasks like **image captioning**. A key advantage of this approach is its **simplified model design and training process**, allowing for **end-to-end learning**. The study highlights that VQA models can be trained **directly from pixels to answers**, eliminating the need for **image region annotations** during preprocessing.

A novel approach is introduced to **VQA** by incorporating **concept-aware representations [30].** Unlike traditional VQA models that rely only on image and text analysis, ConceptBert integrates **external knowledge** from structured sources like **Knowledge Graphs (KGs)** to enhance reasoning capabilities. The model employs a **multi-modal representation** that integrates **concepts, vision, and language,** inspired by the **BERT architecture**. By leveraging **ConceptNet KG**, ConceptBert encodes **common sense knowledge** to improve responses to complex questions that require more than just visual understanding. The approach is evaluated on **OK-VQA and VQA datasets,** demonstrating its effectiveness in handling **knowledge-intensive queries**.

A novel approach to **VQA, Multimodal Residual Networks (MRN)** is introduced which extends deep residual learning [31]. Unlike traditional residual learning, MRN effectively integrates **vision and language** through **element-wise multiplication**, enhancing multimodal representations. The study explores various multimodal architectures and demonstrates **state-of-the-art performance** on VQA datasets for both **Open-Ended and Multiple-Choice tasks**. Additionally, it introduces a method to **visualize attention effects** within joint representations using **backpropagation**, even when spatial information is absent.

More recently, de Faria et al. [32] conducted a comprehensive survey on VQA techniques, outlining the transition from early CNN-RNN architectures to advanced transformer-based and vision-language pre-trained models. Their work highlighted the important role of multimodal fusion and attention mechanisms in improving answer accuracy and generalization across diverse datasets. However, the study primarily focused on architectural trends and lacked in-depth empirical comparisons or critical analysis of model limitations in reasoning for further exploration in practical performance and dataset specific challenges. Their survey also categorized VQA architectures into early fusion, late fusion, and joint embedding strategies, emphasizing how each approach handles multimodal integration. Early fusion methods combined image and text features at the input level, while late fusion techniques processed each modality independently before merging them at the decision stage. Joint embedding models, particularly those based on transformers, enabled deeper cross-modal interactions through attention mechanisms. Despite these advancements, the authors noted that many models still struggle with complex reasoning, compositionality, and external knowledge integration highlighting persistent challenges in achieving human-level understanding in VQA systems.

3. Proposed model

3.1 Methodology

In this research, we introduce a new VQA model based on the architecture proposed in reference [4]. The model receives an image and a natural language question as input. First, the words in the question are converted into numerical vectors using the Skipgram model, and word embeddings are generated. These vectors are then passed through a self-attention mechanism, followed by a Bi-LSTM network for sequential text processing. The output representation from the Bi-LSTM is further refined using a multi-head attention (MHA) module to capture contextual dependencies and generate the final textual representation. Simultaneously, the input image is processed using a 152-layer ResNeXt CNN, which extracts a visual feature vector. This image feature vector is then fed into the same MHA module alongside the textual representation, enabling cross-modal attention and joint feature fusion for answer prediction.

After processing the inputs with neural networks and obtaining their representation vectors, the textual and visual features are concatenated to form a joint representation. Based on Figure 6, this concatenation are performed along the feature dimension, meaning that the output vectors from the Bi-LSTM (text) and the CNN (image) are aligned side-by-side to form a single matrix. This joint matrix representation is then passed through two convolutional layers to extract higher-level fused features, and the resulting output is fed into a Softmax function for answer prediction. Initially, attention weights are computed over the input features, helping to highlight the most relevant components and reveal relationships between modalities.

In other words, multiple attention mechanisms are calculated on the features of the image. The output obtained from this step is then integrated with the features from the inputs using the concatenate method. The final output, a common representation, is passed through two fully connected layers and sent to a Softmax function to predict the final answer and obtain the probabilities for each answer class. Afterward the predicted answer and its probabilities are determined. Multi-head attention mechanism has been simultaneously utilized in both text and image processing. This choice allows the model to identify complex semantic and visual dependencies and establish deeper connections between multimodal data. The use of this mechanism not only enhances textual comprehension in linguistic contexts but also aids in better recognition of key regions in image processing. By leveraging multi-head attention, the model can process multiple levels of information. In text processing, this mechanism enables the preservation of long-range dependencies between words, while in image processing, it facilitates focus on key areas and improves the distribution of visual information. This capability allows the model to generate richer representations of data, ultimately boosting the accuracy of response predictions. The schematic of the proposed visual question answering model in this research is shown in Figure 6.

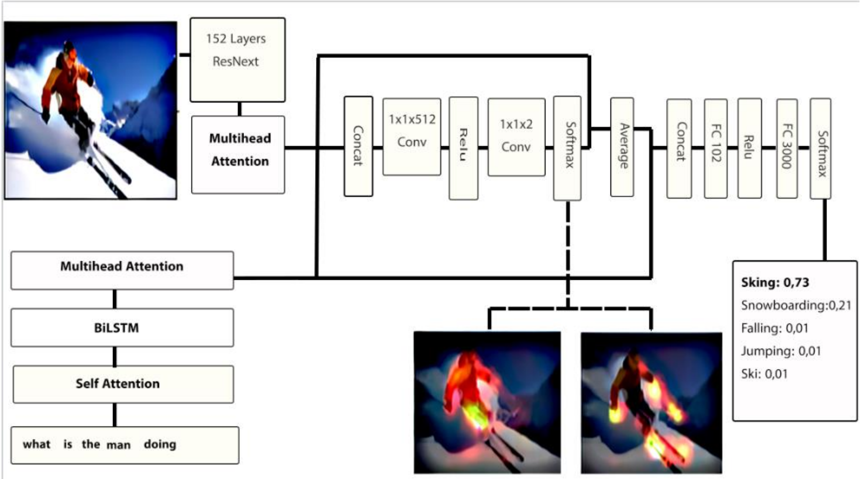

Figure 6: Architecture of the proposed research model

As previously mentioned, the proposed model in this research utilizes a CNN with a ResNeXt architecture of 152 layers to process input images. It should be noted that the reference article [4] uses a CNN with a ResNet architecture, also with 152 layers, for the same purpose. The ResNeXt architecture is an improvement over the ResNet architecture, increasing accuracy and strengthening the object recognition process in images without adding more parameters. Additionally, this architecture reduces error compared to the ResNet architecture. As such, the ResNeXt architecture is one of the best methods for object detection due to its higher accuracy and lower error rate compared to the ResNet architecture.

One of the most important components in a VQA system is the processing of textual input (questions about the image). In the reference article [4], an RNN of the LSTM type is used to process the input text after embedding the words. The input text is given to this network in the form of numerical vectors, from which textual features are extracted to obtain the representation of the text input. In the proposed model of this research, the Skipgram language model is used for text input processing. After embedding, the words are given to an RNN of the BiLSTM type. In fact, in the proposed model, an RNN of the BiLSTM type is used instead of an RNN of the LSTM type. In an RNN, long-term processing is done in one direction, meaning the current input and the output of the past are used in the recurrent network. However, BiLSTM processes information in two directions. In the current input processing, in addition to using the output of the previous processing and past information, future information is also utilized. For this reason, processing in an RNN of the BiLSTM type is better and stronger than processing in an RNN of the simple LSTM type. Therefore, in this research, we changed the input text processing part and instead of using an RNN of the LSTM type, we used an RNN of the BiLSTM type.

In the architecture of the proposed model in the reference article [4], only a soft attention mechanism is used. Meanwhile, in the proposed model of this research, by applying changes to the model from the reference article [4], two additional attention mechanisms are used alongside the soft attention mechanism: the multi-headed attention mechanism and the self-attention mechanism. The reason for using the self-attention mechanism in this research is to better understand and accurately measure the relationships between the words in the input text. Additionally, the multi-headed attention mechanism with eight heads is used to apply eight attention mechanisms to the input simultaneously and in parallel, thus improving input processing.

As mentioned in previous articles, one of the most useful datasets in VQA is the VQA dataset, available in two versions: VQA1.0 and VQA2.0. The reference article [4] utilized both versions of this dataset to train the proposed model. Accordingly, we use both versions of the VQA dataset to train the proposed model.

3.2. Implementation steps and details

As mentioned earlier, this research uses both convolutional networks and RNNs to process textual and image inputs. Initially, the words and questions in both versions of the dataset are tokenized, with 13 words allocated for each question. The words are then embedded using the pre-trained Skipgram language model. The output of this embedding is fed into a self-attention mechanism and then passed to a BiLSTM network with 1024 units. The output of the BiLSTM network is subsequently given to a multi-headed attention mechanism.

For image input, the system resizes images to dimensions of 244×244 pixels before processing them through a 152-layer ResNeXt CNN. The output comes from the fourth convolutional layer. In this implementation, pre-trained weights from ImageNet are used to initialize the ResNeXt model, allowing the network to benefit from prior visual knowledge and accelerate convergence. This architecture employs group normalization, which normalizes feature maps across groups of channels, making it suitable for small batch sizes. A dropout technique with a rate of 0.03 is applied to reduce overfitting. The network trains over eighty epochs, with each batch size set to sixty-four. Additionally, the Adam optimization algorithm is used with a learning rate of 0.0001.

4. Evaluation

In the reference article [4], both versions of the VQA dataset (VQA1.0 and VQA2.0) were used to measure and evaluate the proposed model. Similarly, in this research, we utilize both versions of this dataset to measure and evaluate our proposed visual question and answer model. We allocate 80% of the dataset as training data and the remaining 20% as test data.

In both the reference article [4] and our proposed model, questions are categorized into three types: those with yes or no answers, those with numerical answers, and those with one-word or multi-word answers categorized as other. The frequencies of these question lengths found in both versions of the VQA dataset are illustrated in Figure 7.

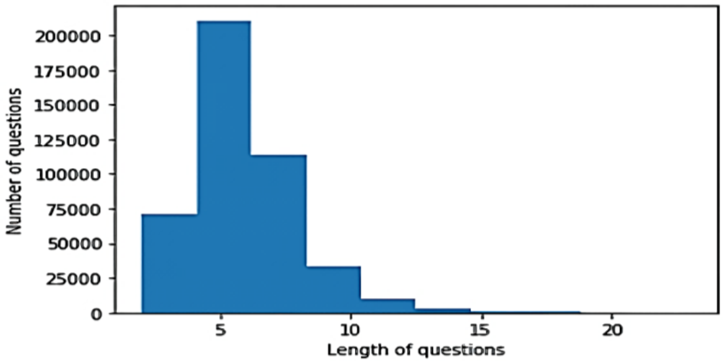

Figure 7: Frequency of question length in the dataset

4.1 Results

The obtained results from the proposed model in this research and the proposed model in the reference article [4] are shown in the diagrams below, based on the two datasets used, VQA1.0 and VQA2.0. The results obtained from evaluating the proposed model indicate a significant improvement over the reference model. Based on conducted experiments, the accuracy of the proposed model on VQA 1.0 dataset has reached to 67.25%, which is 2.65% higher than the 64.6% accuracy of the reference model. Additionally, on the VQA 2.0 dataset, the proposed model achieved 61.57% accuracy demonstrating 1.9% improvement compared to the 59.67% accuracy of the reference model. This increase in accuracy highlights the effectiveness of the proposed architecture in optimizing the integration of visual and textual information. By utilizing multi-head attention in both text and image processing, the model successfully identifies deeper semantic and visual relationships, leading to higher accuracy in feature extraction and response prediction. According to the results, the visual question and answer model proposed in this research demonstrates higher response prediction accuracy than the model proposed in the reference article, based on both datasets. The changes made in this research to the model proposed in the reference article [4] have successfully improved the model.

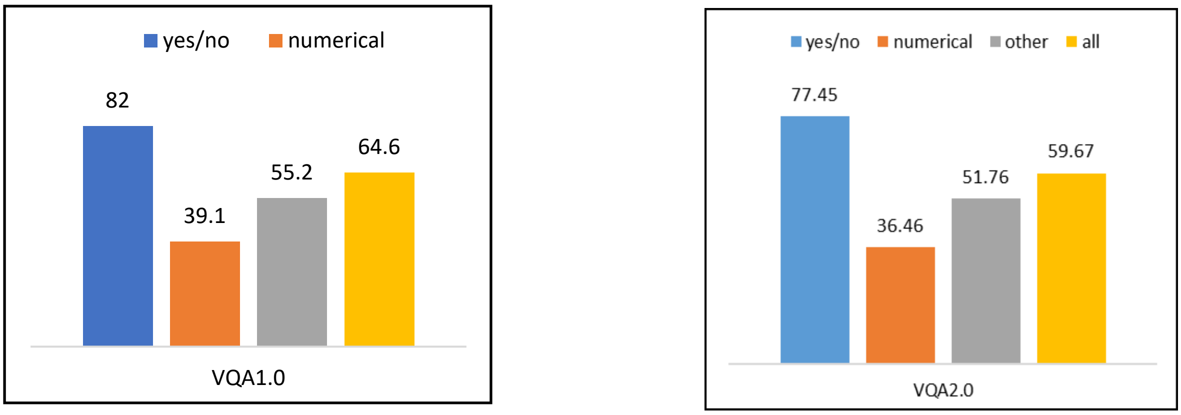

Figure 8: Obtained accuracy in reference article [4]

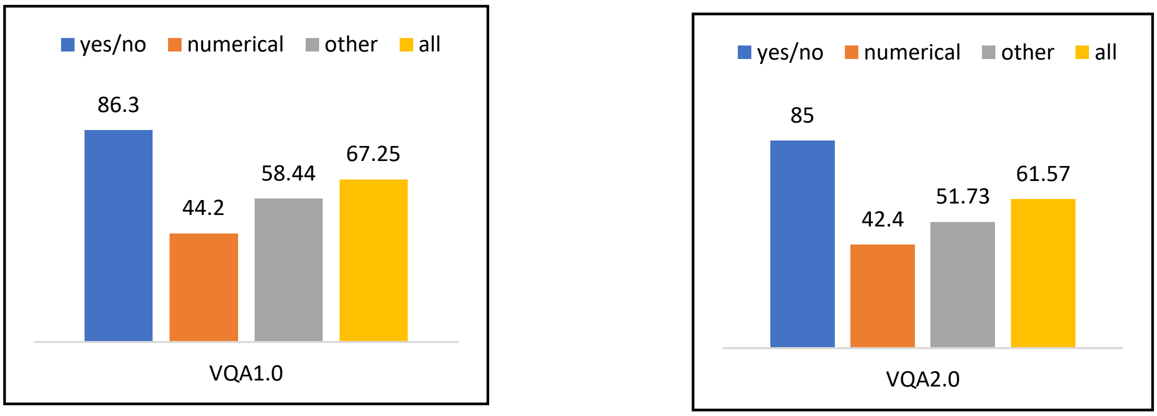

Figure 9: Accuracy obtained by the proposed model

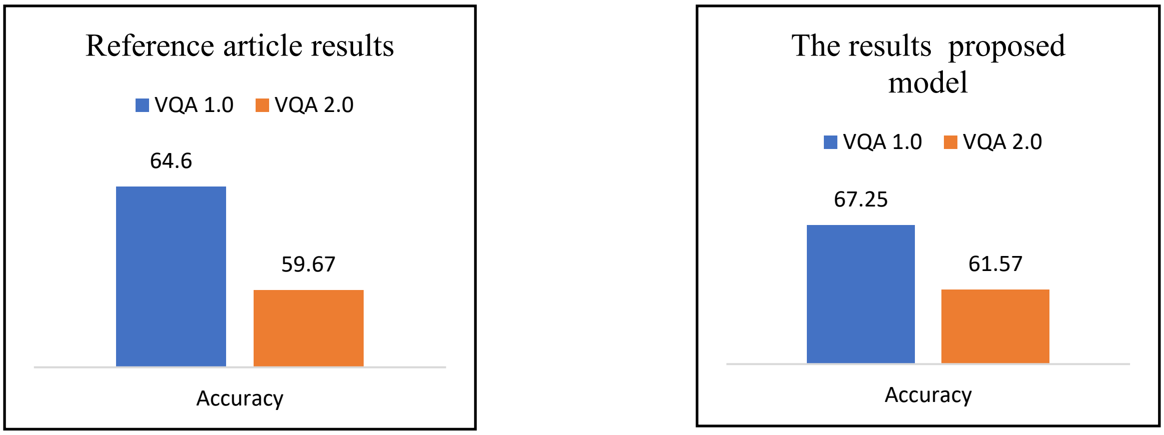

Figure 10: Accuracy of the reference model and the proposed model

The performance of the proposed model in this research is evaluated using both the VQA 1.0 and VQA 2.0 datasets. VQA 1.0 contains over 614,000 questions linked to 204,721 images, while VQA 2.0 includes more than 1.1 million questions, also based on the COCO image dataset. Each question is annotated with 10 human-generated answers, and the questions are categorized into three main types: Yes/No, Number, and Other. The distribution of these categories is approximately 41% Yes/No, 12% Number, and 47% Other. To ensure the reliability of the results, all experiments were repeated five times under identical conditions. We report the average accuracy and standard deviation across these runs to assess the stability of the model's performance. The detailed results of each run are presented in Table 1.

Dataset	Run1	Run2	Run3	Run4	Run5	Average Accuracy (%)	Std. Deviation (%)
VQA 1.0	67.10	67.42	67.33	67.28	67.12	67.25	0.12
VQA 2.0	61.45	61.62	61.70	61.50	61.58	61.57	0.10

Table 1: The detailed results of proposed model during each training run

5. Conclusion

Although a wide range of VQA models have been proposed from early CNN-RNN architectures to attention-based and transformer-driven approaches, there may be some possible limitations. Many existing models struggle with complex reasoning, compositional understanding, and external knowledge integration. Furthermore, most recent models rely heavily on large scale pretraining, which introduces domain-specific biases and limits generalization. The recent survey by de Faria et al. [32] provides a valuable overview of these trends but lacks empirical comparison and critical analysis of model limitations. These gaps highlight the need for more interpretable, efficient, and context-aware architectures. Our proposed model investigated these challenges by integrating self-attention, Bi-LSTM, and multi-head attention mechanisms for richer textual representation, while leveraging deep visual features from ResNeXt and explicitly modeling cross-modal interactions offering more balanced and improved solution for VQA tasks.

In this research, we introduced a new visual question and answer model which an image and a question about the image in natural language are fed as input, and the model must predict an appropriate answer to the question. For this purpose, we used the proposed visual question and answer model from the reference article [4] as the base model, and by making modifications to this model, we introduced a new and more accurate model. We replaced the ResNet convolutional network which contains 152 layers, used in the reference article [4], with the ResNeXt 152 layers for image processing. Additionally, we incorporated two additional attention mechanisms—multi-head attention and self-attention—alongside the attention mechanism used in the base model. These changes led to higher accuracy in response prediction as shown by the results.

6. Future works

Nowadays, we are witnessing the introduction of new architectures of CNN. Researchers in the field of visual question and answer systems are suggested to use newer and stronger CNN architectures such as

Yolo for image processing. It is also advised to use a variety of transformers and larger datasets. The larger the size of the dataset, the better model training process is done, and the accuracy of the predicted final answer is higher.

7. References

[1] J. Andreas, M. Rohrbach, T. Darrell &D. Klein. "Learning to compose neural networks for question answering". In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2016.

[2] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering", in Advances in neural information processing systems, 2015.

[3] 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, pages 2425–2433, 2015.

[4] V. Kazemi, A. Elqursh. "Show, Ask, Attend, and Answer: A Strong Baseline for Visual Question Answering." In: arXiv preprint arXiv: 1704.03162v2, 2017.

[5] X. Zhou, W. Gong, W. Fu, F. Du. "Application of deep learning in object detection". IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS) 2017.

[6] T. Mikolov, I. Sutskever, K. Chen, C. Corrado, S. Greg, J. Dean. "Distributed representations of words and phrases and their compositionality". Advances in Neural Information Processing Systems. arXiv: 1310.4546, 2016.

[7] J. Pennington, R. Socher, and C.D. Manning. "GloVe: Global vectors for word representation." In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

[8] A. Krizhevsky, I. Sutskever, G. Hinton. "ImageNet classification with deep convolutional neural networks" Communications of the ACM. 60 (6): 84–90, 2017.

[9] J. Deng, W. Dong, R. Socher, J. Li, K. Li, and L. Fei-Fei. "Imagenet: A largescale hierarchical image database". In Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2009.

[10] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." In Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2016.

[11] S.xie, R. Gitshick, P. Dollar, Z. Tu and K. He. "Aggregated Residual Transformations for Deep Neural Networks". EEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[12] T. Szandała. "Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks". international Conference on Dependability and Complex Systems, pages 498–505, 2020.

[13] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. "Composing simple image descriptions using web-scale n-grams". In The SIGNLL Conference on Computational Natural Language Learning, 2011.

[14] Y. Kim, "Convolutional neural networks for sentence classification". in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014.

[15] B. Alabassy, M. Safar, M. Watheq EL-Kharashi. "A High-Accuracy Implementation for Softmax Layer in Deep Neural Networks." In Proceedings of the IEEE International Joint Conference on Neural Networks, pages 335–340, 2016.

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," CoRR, vol. abs/1502.03167, 2015.

[17] S. Hochreiter and J. Schmidhuber." Long short-term memory". Neural computation (1735–1780), 1997.

[18] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". Conference on Empirical Methods in Natural Language Processing, 2014.

[19] B. Zhang, H. Wang, L. Jiang, S. Yuan, M. Li. "A Novel Bidirectional LSTM and Attention Mechanism Based Neural Network for Answer Selection in Community Question Answering", Computers, Materials and Continua 61(3): 1273-1288, 2019.

[20] M., Ren, R., Kiros, and R. S., Zemel, "Exploring models and data for image question answering", In Proceedings of the 28th International Conference on Neural Information Processing Systems, Volume 2 (NIPS'15). MIT Press, Cambridge, MA, USA, 2953–2961, 2015.

[21] S., Antol, A., Agrawal, J., Lu, M., Mitchell, D., Batra, C. L., Zitnick, D. Parikh, "VQA: Visual Question Answering", International Conference on Computer Vision (ICCV), 2015.

[22] L. Ma, Z., Lu and H. Li. "Learning to answer questions from image using convolutional neural network." In Proceedings of the AAAI conference on artificial intelligence, vol. 30, no. 1. 2016.

[23] M. Malinowski, M. Rohrbach and M. Fritz, "Ask Your Neurons: A Neural-Based Approach to Answering Questions about Images," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pages 1-9, doi: 10.1109/ICCV.2015.9.

[24] A. S. Toor, H. Wechsler, and M. Nappi, Question action relevance and editing for visual question answering. Multimedia Tools Appl. 78, 3, 2019, 2921–2935.

[25] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam. "Simple baseline for visual question answering." In: arXiv preprint arXiv: 1512.02167, 2015.

[26] A. Jabri, A. Joulin, L.van der Maaten. "Revisiting visual question answering baselines." In: European conference on computer vision. Springer, Cham, 2016.

[27] S. Antol, C. L. Zitnick, and D. Parikh. "Zero-shot learning via visual abstraction". In Proc. Eur. Conf. Comp. Vis, 2014.

[28] J. H. Kim, S. W. Lee, D. Kwak, M. O. Heo, J. Kim, J. W. Ha, and B. T. Zhang. "Multimodal residual learning for visual QA." In Advances in Neural Information Processing Systems 29, 2016.

[29] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen. "In Defense of Grid Features for Visual Question Answering." IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[30] F. Garderes, M. Ziaeefard, B. Abeloos, and F. Lecue. "ConceptBert: Concept-Aware Representation for Visual Question Answering." Findings of the Association for Computational Linguistics: EMNLP 2020.

[31] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. "Deep Compositional Question Answering with Neural Module Networks". In Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2016.

[32] A.C.A.M. de Faria, F.D.C. Bastos, J.V.N.A. da Silva, V.L. Fabris, V.D.S. Uchoa, D.G.D.A. Neto, C.F.G.D. Santos. "Visual Question Answering: A Survey on Techniques and Common Trends in Recent Literature". arXiv preprint arXiv:2305.11033, 2023.