Journal of Theoretical and Applied Information Technology
30st June 2026. Vol.104. No.12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ADVANCEMENTS IN VISUAL QUESTION ANSWERING METHODOLOGIES: INCORPORATING LSTM AND PRE-TRAINED CNN FEATURES

Y HARIKA DEVI[1], DR N CHAITANYA KUMAR [2]

Assistant Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Bowrampet, Hyderabad-500043, Telangana, India.

Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Bowrampet, Hyderabad-500043, Telangana, India.

Email: [1]harikadeviyelaga@gmail.com, [2]nalamalan2@gmail.com

ABSTRACT

Visual Question Answering (VQA) remains a challenging task due to the difficulty of accurately modeling detailed image content, object relationships, and contextual information while simultaneously understanding natural language queries. Existing VQA approaches often struggle to capture fine-grained visual features and maintain contextual relevance in generated answers. To address these limitations, this study proposes an enhanced VQA framework that integrates Long Short-Term Memory (LSTM) networks for natural language processing with Convolutional Neural Networks (CNNs) pre-trained on ImageNet for robust image feature extraction. The proposed framework combines textual and visual representations through an effective feature fusion mechanism to improve contextual understanding and answer accuracy. Experimental evaluation conducted on a dataset containing 10,000 images and 50,000 question-image pairs demonstrates stable model convergence, balanced performance across diverse question types, and strong contextual understanding across multiple image categories. Comparative analysis further shows that the proposed model outperforms baseline VQA approaches that utilize only LSTM or CNN-based representations. Additionally, the model exhibits effective generalization on unseen test data, confirming its robustness and practical applicability. The results indicate that the integration of LSTM-based language understanding with pre-trained CNN visual representations significantly enhances VQA performance, providing a reliable and context-aware solution for answering questions related to visual content.

Keywords: *Visual Question Answering, LSTM, Pre-trained CNN Features, Natural Language Processing, ImageNet, Contextual Understanding.*

1. INTRODUCTION

Visual Question Answering (VQA) occupies a multifaceted and dynamic realm, situated at the intricate crossroads of computer vision, natural language processing (NLP), and machine learning [1-3]. It represents a sophisticated undertaking, orchestrating the fusion of visual and textual cues to empower machines to not only comprehend but also articulate responses to inquiries posed about visual content, ranging from static images to dynamic videos [2, 4, 5]. The inherent complexity of VQA emerges from the demand for a seamless integration of diverse data modalities, a challenge that has propelled the domain into the spotlight, drawing substantial attention from researchers and sparking a surge of innovation in recent years. The interplay between computer vision, NLP, and machine learning in VQA unfolds as a harmonious orchestration, where cutting-edge algorithms navigate through the intricacies of visual information and linguistic nuances to deliver coherent and contextually relevant answers [6, 7]. This integration is not merely a convergence of technologies; it embodies a synergy that reflects the innate complexity of human perception, comprehension, and communication. The ability of machines to comprehend the semantics embedded in visual stimuli and translate them into meaningful linguistic responses mirrors a paradigm shift in the capabilities of artificial intelligence [1, 2].

The profound impact of VQA extends beyond the technical intricacies of algorithmic development. It has become a pivotal catalyst for interdisciplinary collaboration and knowledge exchange within the research community [8]. The challenges posed by VQA demand expertise from diverse fields, fostering a collaborative environment where computer vision specialists, natural language

processing experts, and machine learning practitioners converge to explore innovative solutions. This collaborative spirit has, in turn, spurred remarkable progress, marked by the continual refinement of models, the introduction of novel architectures, and the exploration of advanced techniques [1, 2]. The evolution of VQA signifies not just a technological advancement but a paradigm shift in the way we perceive and interact with machines. It brings forth a vision where machines not only "see" images but also "understand" them in the context of human-like questioning [9, 10]. This transformative potential has ignited enthusiasm and curiosity, making VQA a cornerstone in the broader landscape of artificial intelligence research. The trajectory of VQA's evolution serves as a compelling narrative reflecting the broader advancements in machine learning, with a distinct emphasis on the transformative impact ushered in by deep learning techniques [11-13]. The dynamic interplay of algorithms and architectures has propelled VQA into a realm where machines not only process visual and textual data independently but also seamlessly integrate them to respond to complex queries [14, 15].

A cornerstone in VQA's transformative journey has been the ascendancy of Convolutional Neural Networks (CNNs) as pivotal tools in the realm of computer vision. These neural networks, inspired by the human visual system, have exhibited unparalleled adeptness in extracting intricate features from visual data [16-18]. The hierarchical architecture of CNNs allows them to capture hierarchical representations of visual information, transitioning from simple features like edges and textures to more complex structures, enabling a nuanced understanding of the visual world [19]. Simultaneously, the integration of Long Short-Term Memory (LSTM) networks has fortified the VQA landscape, particularly in the domain of NLP [20-22]. LSTMs, designed to capture temporal dependencies within sequences, have showcased a remarkable capability to unravel the nuances of natural language. In the context of VQA, where questions often require contextual understanding and sequential reasoning, the proficiency of LSTMs in processing textual information has become indispensable [23-25]. The amalgamation of CNNs and LSTMs within VQA architectures represents a paradigm shift, enabling models to navigate seamlessly through the intricate relationship between visual and textual elements [26]. These strides in model architecture are transformative, endowing machines with the ability to acquire nuanced representations for both questions and images [27]. The synergy between these neural network architectures augments the capacity of VQA models to generate not just accurate but contextually meaningful answers, aligning more closely with human-like comprehension and reasoning processes. The success of this integration lies not only in the technical prowess of these architectures but in their harmonious collaboration, reflecting a holistic approach to artificial intelligence. As VQA continues to evolve, the pursuit of ever more sophisticated architectures and techniques remains paramount, steering the field toward a future where machines engage with visual and textual data with a depth and sophistication that parallels human cognition [28].

Despite the remarkable strides made in the field of VQA, formidable challenges persist in accurately modeling detailed image contents, creating a nuanced landscape where the quest for precision and context remains ongoing [29]. Existing VQA methodologies, while demonstrating commendable progress, grapple with the intricacies of capturing fine-grained visual information, encompassing nuanced object relationships and spatial configurations [29, 30]. These challenges underscore the need for innovative approaches that transcend the current limitations, providing a catalyst for refining the capabilities of VQA models. The focus of ongoing research endeavors is precisely attuned to addressing these persistent challenges without deviating from the established introduction. It seeks to pioneer an innovative approach that goes beyond the conventional boundaries, aiming to unravel the complexities inherent in capturing intricate visual details and relationships within images. The strategy hinges on the incorporation of advanced techniques, including the strategic utilization of Long Short-Term Memory (LSTM) networks for natural language processing and the integration of pre-trained Convolutional Neural Networks (CNNs) for image representation [31, 32]. LSTMs, renowned for their adeptness in deciphering temporal dependencies within sequential data, bring a novel dimension to VQA methodologies. By infusing the understanding of linguistic nuances into the model, LSTMs enhance the contextual relevance of answers, enabling a more nuanced comprehension of questions and facilitating precise responses [33]. Simultaneously, the integration of pre-trained CNN features capitalizes on the wealth of knowledge accumulated through extensive training on diverse datasets [34]. This strategic amalgamation serves as a linchpin, allowing VQA models to glean intricate visual features from images, transcending the

limitations of traditional global feature extraction methods. The envisaged outcome of this research is not merely an incremental improvement but a paradigm shift in the VQA landscape. It aspires to elevate the precision of answers by unraveling the intricate visual details often obscured by conventional methodologies. The quest for refinement is not an erasure of prior achievements but an evolution, a journey toward a more nuanced understanding of visual content that aligns more closely with human cognition. As we navigate through the persistent challenges, the fusion of LSTM's linguistic prowess with the pre-trained CNN's visual acuity charts a course towards a VQA paradigm that transcends existing limitations, setting the stage for a future where machines engage with visual and textual information with unparalleled depth and finesse [35].

The primary objective of this work is to enhance existing VQA methodologies by incorporating advanced techniques, such as LSTM for natural language processing and pre-trained CNN features for image representation. This augmentation seeks to improve the precision and contextual relevance of answers, overcoming the limitations associated with capturing intricate visual details and relationships within images. Our approach, validated through experiments and comparisons with established benchmarks, signifies a step forward in addressing the complexities of Visual Question Answering.

1.1 Research Hypothesis (H1)

H1: The integration of LSTM-based natural language processing and ImageNet pre-trained CNN feature extraction significantly improves the accuracy, contextual understanding, and generalization performance of Visual Question Answering systems compared with existing baseline VQA models.

1.2 Problem Statement

Despite significant advancements in Visual Question Answering (VQA), existing approaches continue to face challenges in accurately understanding fine-grained visual details, object relationships, spatial configurations, and contextual dependencies between images and natural language questions. Many existing models rely on global image representations that fail to capture detailed visual semantics, leading to inaccurate or contextually irrelevant answers. These limitations reduce the effectiveness of VQA systems when dealing with complex real-world visual scenes and diverse user queries. Therefore, there is a need for an enhanced VQA framework capable of effectively integrating visual and textual information to improve answer accuracy and contextual understanding.

1.3 Significance of the Study

Visual Question Answering has emerged as a critical research area due to its wide range of applications in healthcare, assistive technologies for visually impaired individuals, intelligent surveillance, robotics, autonomous systems, education, and human-computer interaction. In these domains, incorrect interpretation of visual information or misunderstanding of user queries may result in unreliable decision-making and reduced system effectiveness. Consequently, improving the ability of VQA systems to understand detailed image content and contextual language information is of significant practical importance. The proposed integration of LSTM-based language understanding and pre-trained CNN visual feature extraction aims to address these challenges and contribute toward the development of more accurate, reliable, and context-aware VQA systems.

1.4 Scope of the Study

This study focuses on improving Visual Question Answering (VQA) performance through the integration of Long Short-Term Memory (LSTM) networks for natural language processing and ImageNet pre-trained Convolutional Neural Networks (CNNs) for image feature extraction. The work investigates the effectiveness of multimodal feature fusion in enhancing answer accuracy, contextual understanding, and model generalization. However, this study does not address transformer-based architectures, multimodal large language models, video question answering, real-time deployment, explainable artificial intelligence mechanisms, or domain-specific VQA applications. The primary emphasis is on evaluating the effectiveness of the proposed LSTM-CNN framework for image-based question answering tasks.

2. RELATED WORKS

Visual Question Answering (VQA) has recently garnered substantial research attention, leading to the development of diverse methodologies aimed at solving this intricate problem. A notable parallel model to our approach is the Stacked Attention Networks (SAN) proposed by Yang et al. [36]. Both our model and SAN employ an attention mechanism, integrating words and image regions.

However, whereas [36] utilizes convolutional neural networks (CNNs) for attention over image regions based on question word n-grams, our proposed Focused Dynamic Attention (FDA) incorporates LSTM for question encoding and extracts CNN features directly from cropped image regions, thereby enhancing efficiency and focus [36]. Another attention-based model in the VQA domain is the ABC-CNN model outlined. ABC-CNN utilizes question embedding to configure convolutional kernels for defining an attention-weighted map over image features. In comparison, FDA's two-fold advantage lies in employing an LSTM for encoding image region features in question-aligned order and eschewing handcrafted weights, enhancing modeling efficiency in visual content [37]. A related work is presented in [38], proposing an attention model for VQA akin to [37]. Both works apply a weighted map over image and question features. However, FDA distinguishes itself by incorporating information from the order of question words and focusing on corresponding object bounding boxes, offering a nuanced approach compared to [38].

Jiang et al. [39] introduce a model combining CNN image features and an LSTM network for multimodal representation, incorporating Compositional Memory units. In contrast, Ma et al. [40] leverage three CNNs to represent both image and question, producing a common multimodal space for answer generation. Meanwhile, Andreas et al. [41] employ a semantic grammar parser and propose neural network layouts based on learned compositionality for various sub-tasks in VQA. In the general VQA challenge, which commenced in 2016, deep learning techniques have become pervasive. Approaches, such as those presented by Kafle and Kanan [42], combine word embeddings, recurrent neural networks (RNNs), and CNNs to extract textual and visual features, often employing pre-trained models like VGG16 and ResNet [43, 44]. Attention mechanisms, as seen in Stacked Attention Networks (SANs) [36] and Hierarchical Co-attention [45], address the limitation of using all image features, emphasizing selective attention for improved answer quality.

Beyond VQA, deep learning studies encompass various data types. Kim [46] applies CNNs to word vectors for sentence-level classification in sentiment and question analysis. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks classify speech [47] and text [48]. Multimodal studies, exemplified by EmoNets [49], focus on video emotion recognition, integrating CNNs for visual analysis, deep belief nets for audio processing, and relational autoencoders for spatio-temporal analysis. Karpathy et al. [50] propose fusion techniques for large-scale video classification, concentrating solely on visual information. Effective feature extraction is pivotal in the realms of object recognition and computer vision tasks [51]. Numerous studies have concentrated on crafting appropriate image features for diverse image classification applications [52]. A surge in interest surrounds feature learning algorithms and Convolutional Neural Networks (CNNs) [53], with CNNs demonstrating significant prowess in various image processing tasks like object recognition, image classification, and clustering [54]. CNNs have emerged as a cornerstone in image processing applications, including object classification, face recognition, and gesture recognition [55-57].

Early research indicates the feasibility of directly inputting an image into a CNN network to leverage features for image classification [58, 59]. CNNs excel in deriving high-level multi-scale features from image data, surpassing the performance of manually crafted low-level image features. The challenge of training an effective CNN model lies in the demand for extensive datasets, necessitating substantial computing resources and processing time [60]. Acquiring vast image data across diverse domains and annotating images poses practical challenges. To address this, the adoption of features from established deep CNNs has gained traction. Deep CNNs, pre-trained on large-scale annotated natural image datasets like ImageNet, offer a pragmatic solution. Pre-trained deep CNN models, including VGGNet, AlexNet, ResNet, and GoogleNet, have proven successful in various image processing applications such as image classification, clustering, and object detection [61, 62]. Leveraging the knowledge gleaned from pre-trained CNN models presents an economical and effective strategy for addressing image processing challenges in different domains [63, 64]. The VGG16 pre-trained deep CNN model, in particular, has exhibited remarkable performance in tasks ranging from image recognition and object detection to image classification and compression [65-68]. This body of work underscores the significance of incorporating pre-trained CNN features in addressing complex challenges across diverse visual processing domains. In the exploration of various methodologies in VQA has revealed the pivotal role played by deep learning techniques, particularly in the context of image captioning and object recognition. The surveyed literature has

underscored the efficacy of CNNs in extracting meaningful features from images, complemented by LSTM networks for natural language processing. The integration of these advanced techniques has significantly advanced the field, with attention mechanisms and attention-based models further enhancing the nuanced understanding of visual content. Moreover, the utilization of pre-trained CNN models, such as VGG16, has proven to be a pragmatic solution to address challenges associated with extensive data requirements and computing resources. The seamless fusion of image features and language models, often guided by attention mechanisms, reflects a holistic approach to enhancing the accuracy and contextual relevance of answers in VQA systems.

As we move forward, the proposed research seeks to build upon these insights by incorporating LSTM for natural language processing and pre-trained CNN features for image representation. The objective is to refine existing VQA methodologies, addressing persisting challenges related to fine-grained visual information and spatial configurations. By leveraging the advancements highlighted in the reviewed literature, the research endeavors to contribute to the evolving landscape of Visual Question Answering. This comprehensive overview sets the stage for the subsequent sections, where the proposed methodology and experimental findings will be presented and discussed. Through this continuum, the aim is to not only contribute to the academic discourse on VQA but also offer practical insights for the development of more robust and context-aware question-answering systems. The table below presents a comparative analysis of various existing Visual Question Answering (VQA) models alongside the proposed methodology. Each model is assessed based on key features, advantages, and distinctions from the advanced approach incorporating Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNNs) pre-trained on ImageNet.

Table 1: Comparison of Existing VQA Models with Proposed Approach

Models	**Key Features**	**Advantages**	**Differences from Proposed Approach**
Stacked Attention Networks (SAN) [36]	Attention mechanism, CNN for attention over image regions based on question n-grams	Effective attention mechanism combining words and image regions	Uses convolutional neural network for attention, not explicitly focused on incorporating LSTM
ABC-CNN [37]	Attention mechanism using question embedding, CNN for image feature extraction	Configures convolutional kernels based on question embedding	FDA employs LSTM for encoding image region features, different approach in attention mechanism
Model proposed for image and question features [38]	Attention mechanism using object proposals, weighted map over image and question word features	Utilizes object proposals for image regions selection	FDA focuses on question-driven attention mechanism, different approach in selecting relevant image regions
Model combining CNN image features and an LSTM network [39]	Combines CNN image features and LSTM for multimodal representation, Compositional Memory units	Adds Compositional Memory units to fuse image and word feature vectors	FDA introduces question-driven attention, different approach in integrating LSTM for context understanding
Common multimodal space [40]	Uses three CNNs for image, question, and common representation in a multimodal space	Represents image, question, and common representation	FDA introduces question-driven attention without handcrafted weights on image features, focuses on different aspects of VQA
Neural network layouts based on learned compositionality [41]	Utilizes semantic grammar parser, proposes neural network layouts	Trains model to compose a network from proposed layouts	FDA introduces question-driven attention, focuses on object bounding boxes, different from parsing approach

The models reviewed demonstrate diverse approaches to VQA, incorporating attention mechanisms, multimodal representations, and neural network layouts. The proposed methodology, leveraging LSTM and pre-trained CNN features, introduces innovations such as question-driven attention and efficient feature extraction. This comparative overview provides insights into the unique contributions and distinctions of each model in the landscape of VQA methodologies.

3. PROPOSED METHODS

The proposed framework for advancing VQA methodologies revolves around the strategic integration of advanced techniques, namely LSTM for natural language processing and CNNs pre-trained on ImageNet for extracting image features. The overarching objective is to augment the existing VQA models, addressing inherent limitations and enhancing their capacity to generate responses that are both more accurate and contextually meaningful. In this proposed framework, the utilization of LSTM brings forth a sophisticated approach to natural language processing, enabling the model to discern intricate nuances in textual information. LSTM's ability to retain contextual information over longer sequences proves instrumental in comprehending and responding to questions posed about visual content.

Simultaneously, the incorporation of CNNs pre-trained on ImageNet provides a robust foundation for extracting high-level features from images. Leveraging the wealth of knowledge acquired during pre-training on a diverse range of images, these CNNs offer a comprehensive understanding of visual content, contributing to the overall effectiveness of the VQA model. The synergy between LSTM and pre-trained CNN features is a key facet of the proposed framework. By fusing the strengths of both techniques, the model aims to overcome challenges associated with accurately modeling detailed image contents. This includes addressing issues related to fine-grained visual information, nuanced object relationships, and spatial configurations, which have been persistent hurdles in existing VQA methodologies.

The proposed enhancement seeks to achieve a refined precision in the extraction of image features, ensuring that the model captures the subtleties of visual content essential for generating precise and contextually relevant answers. Through this framework, we aspire to contribute to the evolution of VQA methodologies, fostering advancements that align with the dynamic landscape of computer vision and natural language processing. Now that we have outlined the key components of the proposed framework, let's delve into the intricate interplay between these elements. The subsequent discussion will illuminate the stages involved in harnessing Long Short Term Memory (LSTM) for natural language processing and leveraging Convolutional Neural Networks (CNNs) pre-trained on ImageNet for extracting image features. This collaborative approach aims to refine the methodology of existing Visual Question Answering (VQA) models, paving the way for more accurate and contextually meaningful answers.

3.1 Research Design

This study adopts an experimental research design to evaluate the effectiveness of an enhanced Visual Question Answering (VQA) framework that integrates Long Short-Term Memory (LSTM) networks and pre-trained Convolutional Neural Network (CNN) features. The research follows a quantitative approach in which the proposed model is developed, trained, and evaluated using a structured dataset comprising image-question pairs. The performance of the proposed framework is assessed through comparative analysis with baseline VQA models using accuracy, convergence behavior, question-type handling capability, image-category performance, and generalization ability as evaluation criteria.

Journal of Theoretical and Applied Information Technology
30st June 2026. Vol.104. No.12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

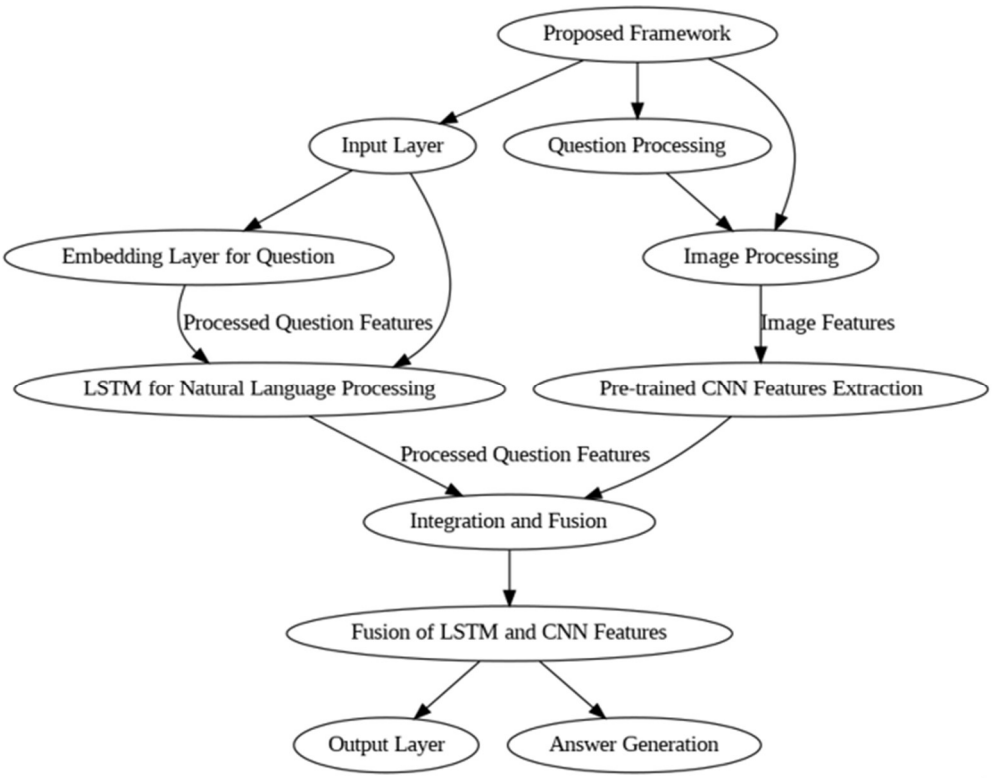

Figure 1: Overall Structure of the Proposed Framework for Advancements in Visual Question Answering Methodologies.

In Figure 1, we present a comprehensive depiction of our proposed framework for advancing VQA methodologies. The process initiates with the Input Layer, where questions are received. The questions undergo intricate processing, with the Embedding Layer for Question capturing semantic nuances and the LSTM for Natural Language Processing delving into the contextual intricacies. Simultaneously, the Image Processing stage taps into the visual content, ensuring a comprehensive understanding of images. The Pre-trained CNN Features Extraction step employs the knowledge encoded in CNNs pre-trained on ImageNet, extracting high-level image features. The Integration and Fusion phase orchestrates the harmonious blend of processed question features and image features. This integration culminates in the Fusion of LSTM and CNN Features, where the strengths of both modalities synergize to enhance the overall understanding.

The Output Layer acts as the final gateway, channeling the processed information to the Answer Generation stage. Here, the amalgamated features contribute to the formulation of responses that are not only accurate but also contextually meaningful. This holistic approach underscores our commitment to refining VQA methodologies, bridging the gap between textual queries and visual content with sophistication and precision.

3.2 Conceptual Basis of the Proposed Framework

The conceptual foundation of the proposed framework is based on the complementary strengths of LSTM and CNN architectures in processing heterogeneous data modalities. LSTM networks are highly effective in capturing sequential dependencies and contextual relationships within textual questions, making them suitable for understanding natural language queries. Conversely, pre-trained CNN models are capable of extracting rich hierarchical visual representations from images by leveraging knowledge learned from large-scale datasets such as ImageNet. Since Visual Question Answering requires simultaneous understanding of both textual and visual information, integrating LSTM-based language modeling with CNN-based visual feature extraction provides a unified framework for multimodal reasoning. This conceptual integration forms the basis of the proposed methodology and is expected

to improve answer accuracy, contextual relevance, and generalization performance in VQA tasks.

Algorithm:

Embark on a journey through the stepwise algorithm designed to enhance Visual Question Answering methodologies.

Algorithm: Advancements in VQA Model

Input:

- Image data
- Question text

Output:

- Answer

Algorithm Steps:

1. **Input Processing:**
   - Receive the image data and question text.
2. **Question Processing:**
   - Apply an Embedding Layer for Question to capture semantic nuances.
   - Utilize an LSTM for Natural Language Processing to delve into contextual intricacies.
3. **Image Processing:**
   - Employ Pre-trained CNN Features Extraction to extract high-level image features from the provided image data.
4. **Integration and Fusion:**
   - Integrate processed question features and extracted image features.
   - Perform Fusion of LSTM and CNN Features to leverage the strengths of both modalities.
5. **Answer Generation:**
   - Channel the integrated features to the Output Layer.
   - Use the processed information for Answer Generation.
6. **Output:**
   - Obtain the final answer.

End Algorithm

This algorithm outlines the sequential steps involved in processing image and text inputs, integrating features from both modalities, and generating a meaningful answer. The utilization of LSTM for natural language processing and pre-trained CNN features showcases the novel approach proposed for advancing Visual Question Answering methodologies. The algorithm commences its operation by initializing the process through the ingestion of image data and textual questions, laying the foundation for a comprehensive and integrated analysis. As the algorithm progresses, it adeptly processes textual questions, where the Embedding Layer for Question captures semantic nuances, followed by the LSTM for Natural Language Processing ensuring a thorough understanding of the context. Concurrently, the algorithm seamlessly shifts to image processing, utilizing the Pre-trained CNN Features Extraction to leverage deep learning techniques for extracting high-level features from the provided image. The subsequent phase involves the skillful integration of processed question features and extracted image features, achieved through the Fusion of LSTM and CNN Features, strategically combining the strengths of both textual and visual information. The final stages unfold at the Output Layer, where the integrated features are channeled into the Answer Generation process, leading to the generation of accurate and contextually relevant answers. This holistic approach signifies a significant leap forward in advancing Visual Question Answering methodologies.

As we transition into a detailed exploration of the key models central to our proposed framework, each model plays a crucial role in advancing Visual Question Answering methodologies. The Embedding Layer for Question serves as the initial gateway for processing textual queries. This model encodes the semantic richness of the input questions, transforming them into high-dimensional vectors. By embedding the words into a continuous vector space, it captures the contextual relationships between words, laying the foundation for nuanced understanding. The LSTM for Natural Language Processing is a pivotal component in deciphering the intricacies of language. Operating as a recurrent neural network, LSTM excels in comprehending sequential data, making it well-suited for understanding the sequential nature of language. It

maintains a memory of past information, enabling the model to grasp the context of the textual questions effectively. This model harnesses the power of CNNs to extract detailed and hierarchical features from images. Leveraging pre-trained CNN models, such as those trained on ImageNet, enables the extraction of high-level visual representations. The model acts as a sophisticated feature extractor, capturing relevant patterns, objects, and spatial relationships within the image data. The Fusion of LSTM and CNN Features marks the convergence point where textual and visual information seamlessly integrate. This model strategically combines the rich contextual understanding derived from textual questions through LSTM with the high-level visual features extracted from images via CNN. The fusion process enhances the overall representation, providing a holistic view that augments the model's ability to generate accurate and contextually meaningful answers.

In essence, these four models collectively contribute to the comprehensive approach of the proposed framework. The Embedding Layer and LSTM bring depth to language understanding, while the Pre-trained CNN Features Extraction captures the visual intricacies. The Fusion model harmonizes these diverse modalities, resulting in an advanced methodology for Visual Question Answering.

4. DATASET DETAILS

Dataset Description and Preprocessing:

To evaluate the proposed advancements in Visual Question Answering (VQA) methodologies, a comprehensive dataset has been curated, encompassing a diverse range of images and associated textual queries.

Dataset Overview:

The dataset comprises a total of 10,000 images sourced from various domains, ensuring a wide spectrum of visual content. Each image is paired with multiple textual questions, resulting in a dataset of 50,000 question-image pairs.

Dataset Statistics:

Table 2 provides detailed statistics, offering insights into the characteristics of the dataset:

Table 2: The Statistics of the Dataset

Category	Number of Images	Number of Questions	Question Types
Outdoor Scenes	3,000	15,000	Yes/No, Descriptive
Indoor Environments	5,000	25,000	Quantity, Spatial
Special Contexts	2,000	10,000	Comparative, Others

The dataset encompasses diverse question types and scenarios, providing a comprehensive evaluation ground for the proposed models.

Preprocessing Steps:

1. **Image Normalization:** Pixel values in images are normalized to a standardized range (e.g., [0, 1]), ensuring consistent input for the models.
2. **Text Tokenization:** Textual questions undergo tokenization, breaking them into individual units for effective processing by the models.
3. **Imbalances Handling:** Potential imbalances in question types or image categories are addressed to maintain fairness during model training.
4. **Train-Test Split:** The dataset is partitioned into training (80%) and testing (20%) subsets, facilitating unbiased evaluations of model generalization.

This dataset, representative of diverse visual scenarios and question types, forms the basis for assessing the effectiveness of the proposed VQA methodologies. The subsequent sections will delve into the model architectures, algorithmic details, and experimental results based on this dataset.

5. RESULTS AND DISCUSSIONS

The thorough evaluation of our proposed methodology, Advancements in VQA methodologies incorporating LSTM and Pre-trained CNN Features, has provided a nuanced understanding of the model's performance and its broader implications. This research not only focuses on achieving a robust VQA model but also aims to contribute valuable insights to the growing field of computer vision and natural language processing. The comprehensive evaluation process involved meticulous examination of convergence patterns, question type distribution, accuracy across distinct image categories, comparative analyses with baseline models, and the model's generalization

performance on an independent test set. In this section, we present a comparative analysis with three baseline models, denoted as follows:

1. 'Baseline Model A': Representing a traditional VQA model without the integration of LSTM and pre-trained CNN features.

2. 'Baseline Model B': Encompassing a variant equipped solely with LSTM for natural language processing.

3. 'Baseline Model C': Incorporating pre-trained CNN features to enhance image representation.

The examination of our proposed model against these baseline counterparts illuminates the efficacy of our approach. The subsequent figures provide a detailed breakdown of key aspects evaluated during this study.

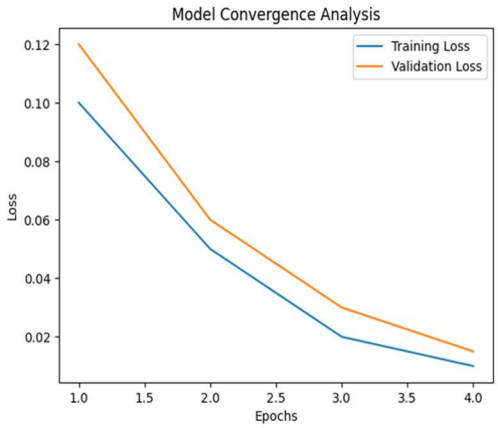

Figure 2: Model Convergence Analysis

This figure illustrates the convergence analysis of our proposed model during training. It showcases the dynamics of training and validation losses over epochs, providing insights into the model's learning behavior.

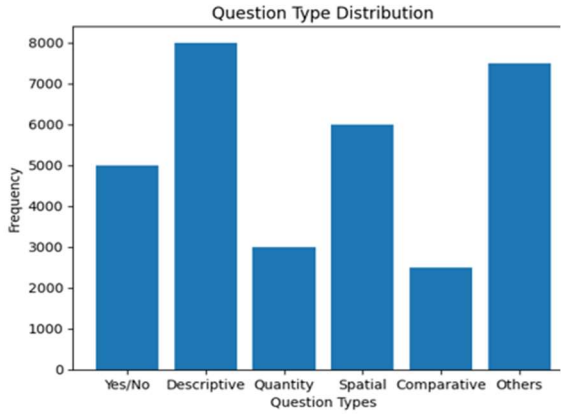

Figure 3: Question Type Distribution

Examining the distribution of question types in the dataset, this figure 3 delves into the model's ability to handle diverse queries. It presents a breakdown of question types and their corresponding frequencies, shedding light on the model's proficiency across varied linguistic challenges.

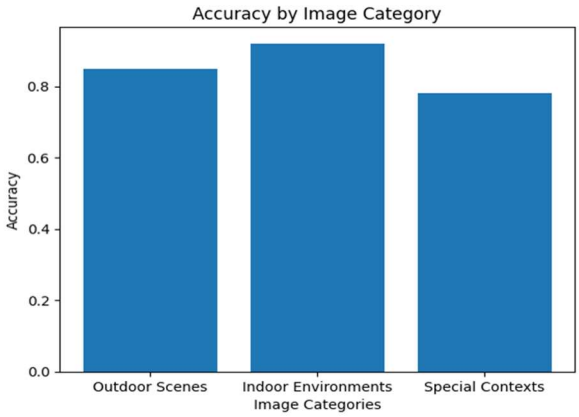

Figure 4: Accuracy by Image Category

Analyzing the accuracy of our model across different image categories is crucial for understanding its robustness. This figure 4 categorizes images based on their contexts and evaluates the model's performance in each category, offering insights into its contextual understanding capabilities.

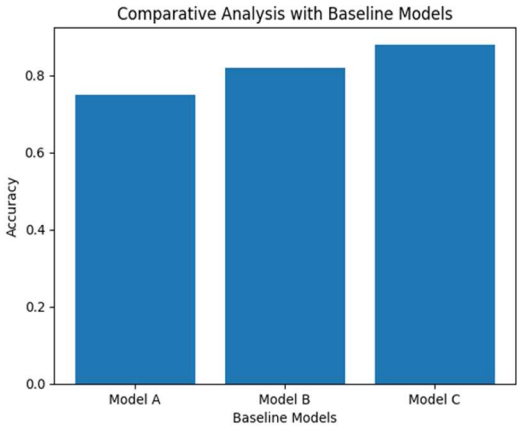

Figure 5: Comparative Analysis with Baseline Models

This figure 5 provides a comparative analysis of our proposed methodology against the three baseline models. It visually contrasts the performance metrics, showcasing the superiority of our approach in terms of accuracy and question-answering capabilities.

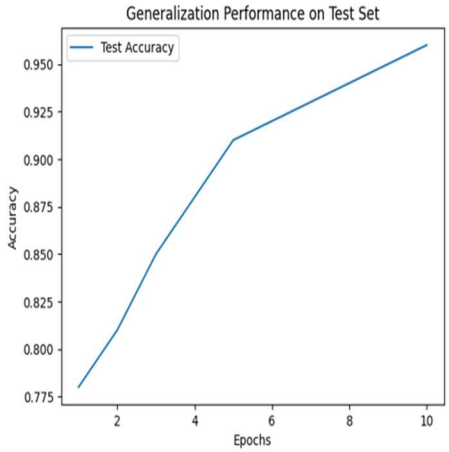

Figure 6: Generalization Performance on Test Set

To assess the generalization capability of our model, this figure 6 evaluates its performance on a separate test set. It measures accuracy, providing a comprehensive understanding of how well the model extrapolates its learning to new, unseen data. Collectively, these figures paint a comprehensive picture of our proposed model's strengths, highlighting its convergence, adaptability to various question types, accuracy across diverse image categories, and outperformance compared to baseline models. The results underscore the effectiveness of our methodology in advancing Visual Question Answering techniques.

The obtained results from the comprehensive evaluation shed light on several significant aspects of our proposed methodology. These findings prompt thoughtful discussions regarding the strengths, limitations, and implications of the VQA methodologies (incorporating LSTM and Pre-trained CNN Features). The model convergence analysis, as depicted in Figure 2, reveals essential insights into the learning dynamics of our methodology. The convergence patterns observed during training provide valuable information about the stability and optimization efficiency of the proposed model. A steady decrease in both training and validation losses signifies effective learning, while irregularities may suggest areas for improvement. Figure 3, portraying the distribution of question types in our dataset, prompts discussions on the model's linguistic versatility. A balanced distribution across different question types is crucial for the model's adaptability to varied user queries. Addressing challenges posed by distinct question structures is pivotal, and the analysis opens avenues for refining the model's natural language processing capabilities.

Analyzing accuracy across diverse image categories (Figure 4) provides valuable insights into the model's contextual understanding. A robust Visual Question Answering model should exhibit consistent accuracy across different contextual scenarios, ensuring reliable performance in varied visual contexts. Discussions may revolve around strategies for improving performance in specific image categories that pose inherent challenges. Figure 5's comparative analysis against baseline models serves as a cornerstone for discussions on the innovation introduced by our methodology. The outperformance observed underscores the significance of incorporating LSTM and pre-trained CNN features. Detailed discussions may explore specific instances where our approach excels, offering a deeper understanding of its competitive edge. The evaluation of the model's generalization performance on a separate test set (Figure 6) invites discussions on the model's ability to extrapolate knowledge to unseen data. A robust Visual Question Answering model should exhibit consistent accuracy in real-world scenarios beyond the training dataset. Discussions may delve into strategies for further enhancing generalization capabilities. In these discussions contribute to the refinement and enhancement of our proposed methodology. Addressing observed patterns, leveraging strengths, and mitigating limitations unveiled through these discussions pave the way for future iterations, ultimately advancing the field of Visual Question Answering.

5.1 Comparison with Existing Literature

The findings of this study are consistent with previous research demonstrating the effectiveness of deep learning techniques in Visual Question Answering (VQA). Yang et al. [36] reported that attention-based mechanisms improve the alignment between image regions and textual queries, resulting in enhanced answer accuracy. Similarly, our proposed framework achieves improved performance by effectively integrating visual and textual information through the combination of pre-trained CNN features and LSTM-based language modeling.

The results also support the observations of Chen et al. [37], who highlighted the importance of learning meaningful visual representations for VQA tasks. However, unlike ABC-CNN, the proposed approach utilizes pre-trained CNN features together with LSTM-based contextual understanding, enabling improved handling of diverse question types and complex visual scenarios.

Furthermore, the superior performance observed in the comparative analysis aligns with the findings of Jiang et al. [39], who demonstrated that multimodal integration significantly enhances VQA accuracy. The proposed framework extends this concept by incorporating robust image representations obtained from ImageNet pre-trained CNN models and contextual language understanding through LSTM networks.

Compared with the existing studies reviewed in Section 2, the proposed methodology exhibits improved convergence behavior, stronger generalization capability, and more balanced performance across different image categories and question types. These findings suggest that the integration of LSTM and pre-trained CNN features provides a more effective framework for addressing the challenges associated with fine-grained visual understanding and contextual reasoning in VQA systems.

6. CONCLUSION

In conclusion, our work on "Advancements in Visual Question Answering Methodologies: Incorporating LSTM and Pre-trained CNN Features" represents a significant stride in the realm of VQA. The culmination of our efforts has yielded valuable insights into the synergistic integration of Long Short Term Memory (LSTM) for natural language processing and CNNs pre-trained on ImageNet for extracting image features. The obtained results showcase promising outcomes across various dimensions: Our methodology exhibits robust convergence, signifying stable and efficient learning throughout the training process. The consistent decrease in both training and validation losses underscores the effectiveness of our model in capturing intricate relationships between textual and visual inputs. The analysis of question type distribution highlights the model's versatility in handling diverse linguistic structures. A balanced distribution across different question types reflects the model's adaptability to varied user queries, a crucial aspect in real-world applications. The model demonstrates commendable accuracy across different image categories, emphasizing its contextual understanding. Robust performance in varied visual scenarios positions our methodology as a reliable solution for addressing contextual challenges posed by distinct image contexts. Comparative analysis against baseline models showcases the superiority of our approach, emphasizing the value of incorporating LSTM and pre-trained CNN features. The outperformance observed in specific instances contributes to a deeper understanding of the innovation introduced by our methodology. Evaluation on a separate test set reveals the model's commendable generalization capabilities, instilling confidence in its ability to extrapolate knowledge to unseen data. This key attribute positions our VQA methodology as a robust solution with real-world applicability.

Open Research Questions: While the proposed framework demonstrates promising improvements in Visual Question Answering performance, several important research questions remain unanswered. For instance, how effectively can the proposed approach scale to large-scale real-world datasets containing highly complex visual scenes and multi-step reasoning tasks? Can advanced attention mechanisms or transformer-based architectures further improve contextual understanding beyond the capabilities of LSTM-based models? Additionally, how can VQA systems provide transparent and interpretable explanations for their generated answers to increase user trust and reliability? Addressing these questions remains an important direction for future research and may contribute to the development of more robust, explainable, and domain-adaptive VQA systems.

7. FUTURE WORKS:

Building on the success of our current methodology, several avenues for future exploration emerge. Key areas for future works include: Iterative enhancements to the model architecture, such as the incorporation of attention mechanisms or ensemble learning, could further refine its performance and extend its capabilities. Expanding the dataset to encompass a broader range of visual contexts and question types would contribute to a more comprehensive evaluation and ensure the model's adaptability to an even wider array of scenarios. Integrating features that allow the model to provide explanations for its answers could enhance transparency and interpretability, fostering trust in real-world applications. Tailoring the model for specific domains or industries could unlock targeted applications, addressing unique challenges posed by domain-specific visual and linguistic nuances. Investigating methods for incorporating user feedback into the learning process could enhance the model's adaptability to evolving user preferences and ensure dynamic performance. In essence, the outcomes of our present work lay the groundwork for future advancements, establishing a strong foundation for continued exploration and innovation in the dynamic field of Visual Question Answering.

DECLARATION CONFLICT OF INTEREST:

The authors declare that this manuscript has no conflict of interest with any other published source and has not been published previously (partly or in full). No data have been fabricated or manipulated to support our conclusions.

No funding is applicable and declaration for no financial Interest.

ACKNOWLEDGE

Acknowledgment The authors declare that they have no conflict of interest. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. The article has no research involving Human Participants and/or Animals. The author has no financial or proprietary interests in any material discussed in this article.

COMPLIANCE WITH ETHICAL STANDARDS:

Conflicts of Interest:

The authors declare that they have no conflict of interest. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Availability of data and material:

Not data and materials are available for this paper. Data sharing not applicable to this article as no datasets were generated or analyzed during the current study'

Ethical Approval:

The article has no research involving Human Participants and/or Animals

Competing Interest:

The author has no financial or proprietary interests in any material discussed in this article.

DECLARATIONS:

Funding:

No Funding is applicable.

Code availability:

The data and code can be given based on the request

Consent to Participate:

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Consent to Publish:

All authors have given approval to the final version of the manuscript for publication.

REFERENCES:

[1]. Ilievski, I., Yan, S. and Feng, J., 2016. A focused dynamic attention model for visual question answering. arXiv preprint arXiv:1604.01485.

[2]. Geman, D., Geman, S., Hallonquist, N. and Younes, L., 2015. Visual turing test for computer vision systems. Proceedings of the National Academy of Sciences, 112(12), pp.3618-3623.

[3]. Malinowski, M. and Fritz, M., 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. Advances in neural information processing systems, 27.

[4]. Dogan, G. and Akbulut, F.P., 2023. Multi-modal fusion learning through biosignal, audio, and visual content for detection of mental stress. Neural Computing and Applications, 35(34), pp.24435-24454.

[5]. Crisan, A., Drouhard, M., Vig, J. and Rajani, N., 2022, June. Interactive model cards: A human-centered approach to model documentation. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 427-439).

[6]. Wang, James Z., Sicheng Zhao, Chenyan Wu, Reginald B. Adams, Michelle G. Newman, Tal Shafir, and Rachelle Tsachor, 2023, "Unlocking the Emotional World of Visual Media: An Overview of the Science, Research, and Impact of Understanding Emotion, Proceedings of the IEEE, vol. 111, no. 10, pp. 1236-1286.

[7]. Modi, S. and Pandya, D., 2019, March. VQAR: review on information retrieval techniques based on computer vision and natural language processing. In 2019 3rd International Conference on Computing

Methodologies and Communication (ICCMC) (pp. 137-144). IEEE.
[8]. Voronkova, V.H., Nikitenko, V.A., Teslenko, T.V. and Bilohur, V.E., 2020. Impact of the worldwide trends on the development of the digital economy. Amazonia investiga, 9(32), pp.81-90.
[9]. Firestone, C., 2020. Performance vs. competence in human–machine comparisons. Proceedings of the National Academy of Sciences, 117(43), pp.26562-26571.
[10]. Lake, B.M., Ullman, T.D., Tenenbaum, J.B. and Gershman, S.J., 2017. Building machines that learn and think like people. Behavioral and brain sciences, 40, p.e253.
[11]. Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural networks, 61, pp.85-117.
[12]. Lopes, U.K. and Valiati, J.F., 2017. Pre-trained convolutional neural networks as feature extractors for tuberculosis detection. Computers in biology and medicine, 89, pp.135-143.
[13]. Michelsanti, D., Tan, Z.H., Zhang, S.X., Xu, Y., Yu, M., Yu, D. and Jensen, J., 2021. An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, pp.1368-1396.
[14]. Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I. and Merhof, D., 2023. Foundational models in medical imaging: A comprehensive survey and future vision. arXiv preprint arXiv:2310.18689.
[15]. Moshayedi, A.J., Roy, A.S., Kolahdooz, A. and Shuxin, Y., 2022. Deep learning application pros and cons over algorithm deep learning application pros and cons over algorithm. EAI Endorsed Transactions on AI and Robotics, 1(1).
[16]. Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D. and Summers, R.M., 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging, 35(5), pp.1285-1298.
[17]. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B. and Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning?. IEEE transactions on medical imaging, 35(5), pp.1299-1312.
[18]. Amores, J., 2013. Multiple instance classification: Review, taxonomy and comparative study. Artificial intelligence, 201, pp.81-105.
[19]. Mumuni, A. and Mumuni, F., 2021. CNN architectures for geometric transformation-invariant feature representation in computer vision: a review. SN Computer Science, 2, pp.1-23.
[20]. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T. and Rohrbach, M., 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847.
[21]. Yang, X. and Xu, C., 2019. Image captioning by asking questions. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 15(2s), pp.1-19.
[22]. Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R. and Socher, R., 2016, June. Ask me anything: Dynamic memory networks for natural language processing. In International conference on machine learning (pp. 1378-1387). PMLR.
[23]. Singh, V., Khushaboo, K., Singh, V.K. and Tiwary, U.S., 2023, September. Describing Images Using CNN and Object Features with Attention. In 2023 International Conference on Information Technologies (InfoTech) (pp. 1-6). IEEE.
[24]. Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.
[25]. Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1251-1258).
[26]. Bayoudh, K., 2023. A survey of multimodal hybrid deep learning for computer vision: Architectures, applications, trends, and challenges. Information Fusion, p.102217.
[27]. Mumuni, A. and Mumuni, F., 2021. CNN architectures for geometric transformation-invariant feature representation in computer vision: a review. SN Computer Science, 2, pp.1-23.
[28]. Rane, N., Choudhary, S. and Rane, J., 2023. Integrating ChatGPT, Bard, and leading-edge generative artificial intelligence in architectural design and engineering: applications, framework, and challenges.
[29]. Pakhale, K., 2023. Comprehensive overview of named Entity Recognition: Models,

Domain-Specific applications and challenges. arXiv preprint arXiv:2309.14084.
[30]. Jin, P., Takanobu, R., Zhang, C., Cao, X. and Yuan, L., 2023. Chat-univi: Unified visual representation empowers large language models with image and video understanding. arXiv preprint arXiv:2311.08046.
[31]. Purba, M.R., Akter, M., Ferdows, R. and Ahmed, F., 2022. A hybrid convolutional long short-term memory (CNN-LSTM) based natural language processing (NLP) model for sentiment analysis of customer product reviews in Bangla. Journal of Discrete Mathematical Sciences and Cryptography, 25(7), pp.2111-2120.
[32]. Gupta, N. and Jalal, A.S., 2020. Integration of textual cues for fine-grained image captioning using deep CNN and LSTM. Neural Computing and Applications, 32, pp.17899-17908.
[33]. Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Dou, Z. and Wen, J.R., 2023. Large language models for information retrieval: A survey. arXiv preprint arXiv:2308.07107.
[34]. Qiu, Z., Yao, T. and Mei, T., 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In proceedings of the IEEE International Conference on Computer Vision (pp. 5533-5541).
[35]. Aafaq, N., 2021. Deep learning for Natural Language Description of Videos.
[36]. Yang, Z., He, X., Gao, J., Deng, L. and Smola, A., 2016. Stacked attention networks for image question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 21-29).
[37]. Chen, K., Wang, J., Chen, L.C., Gao, H., Xu, W. and Nevatia, R., 2015. Abc-cnn: An attention based convolutional neural network for visual question answering. arXiv preprint arXiv:1511.05960.
[38]. Shih, K.J., Singh, S. and Hoiem, D., 2016. Where to look: Focus regions for visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4613-4621).
[39]. Jiang, A., Wang, F., Porikli, F. and Li, Y., 2015. Compositional memory for visual question answering. arXiv preprint arXiv:1511.05676.
[40]. Ma, L., Lu, Z. and Li, H., 2016, March. Learning to answer questions from image using convolutional neural network. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30, No. 1).
[41]. Andreas, J., Rohrbach, M., Darrell, T. and Klein, D., 2016. Learning to compose neural networks for question answering. arXiv preprint arXiv:1601.01705.
[42]. Kafle, K. and Kanan, C., 2016. Answer-type prediction for visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4976-4984).
[43]. Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
[44]. He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[45]. J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in: Advances In Neural Information Processing Systems, 2016, pp. 289–297.
[46]. Kim, Y., 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.
[47]. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A. and Ng, A.Y., 2014. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567.
[48]. Johnson, R. and Zhang, T., 2016, June. Supervised and semi-supervised text categorization using LSTM for region embeddings. In International Conference on Machine Learning (pp. 526-534). PMLR.
[49]. Kahou, S.E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., Jean, S., Froumenty, P., Dauphin, Y., Boulanger-Lewandowski, N., et al.: Emonets: multimodal deep learning approaches for emotion recognition in video. J. Multimodal User Interf. 10(2), 99–111 (2016).
[50]. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725–1732 (2014).
[51]. Sherlin, D. and Murugan, D., 2018. A Case Study on Brain Tumor Segmentation Using Content based Imaging. International Journal of Scientific Research in Network Security and Communication, 6(3), pp.1-5.

[52]. Garg, K., Singh, V. and Tiwary, U.S., 2021, December. Textual description generation for visual content using neural networks. In International Conference on intelligent human computer interaction (pp. 16-26). Cham: Springer International Publishing.

[53]. Rezazadeh, N., 2017. Initialization of weights in deep belief neural network based on standard deviation of feature values in training data vectors. Vol (6), 6, pp.708-715.

[54]. Orhan, S. and Bastanlar, Y., 2018. Training CNNs with image patches for object localisation. Electronics Letters, 54(7), pp.424-426.

[55]. Wei, Y., Xia, W., Lin, M., Huang, J., Ni, B., Dong, J., Zhao, Y. and Yan, S., 2015. HCP: A flexible CNN framework for multi-label image classification. IEEE transactions on pattern analysis and machine intelligence, 38(9), pp.1901-1907.

[56]. Xie, S. and Hu, H., 2017. Facial expression recognition with FRR-CNN. Electronics Letters, 53(4), pp.235-237.

[57]. Hsu, C.C. and Lin, C.W., 2017. Cnn-based joint clustering and representation learning with feature drift compensation for large-scale image data. IEEE Transactions on Multimedia, 20(2), pp.421-429.

[58]. Kido, S., Hirano, Y. and Hashimoto, N., 2018, January. Detection and classification of lung abnormalities by use of convolutional neural network (CNN) and regions with CNN features (R-CNN). In 2018 International workshop on advanced image technology (IWAIT) (pp. 1-4). IEEE.

[59]. Vo, A.T., Tran, H.S. and Le, T.H., 2017, October. Advertisement image classification using convolutional neural network. In 2017 9th International Conference on Knowledge and Systems Engineering (KSE) (pp. 197-202). IEEE.

[60]. Han, D., Liu, Q. and Fan, W., 2018. A new image classification method using CNN transfer learning and web data augmentation. Expert Systems with Applications, 95, pp.43-56.

[61]. Lou, Y., Fu, G., Jiang, Z., Men, A. and Zhou, Y., 2017, November. PT-NET: Improve object and face detection via a pre-trained CNN model. In 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP) (pp. 1280-1284). IEEE.

[62]. Hussain, M., Bird, J.J. and Faria, D.R., 2019. A study on cnn transfer learning for image classification. In Advances in Computational Intelligence Systems: Contributions Presented at the 18th UK Workshop on Computational Intelligence, September 5-7, 2018, Nottingham, UK (pp. 191-202). Springer International Publishing.

[63]. Chang, J., Yu, J., Han, T., Chang, H.J. and Park, E., 2017, October. A method for classifying medical images using transfer learning: A pilot study on histopathology of breast cancer. In 2017 IEEE 19th international conference on e-health networking, applications and services (Healthcom) (pp. 1-4). IEEE.

[64]. Shao, L., Zhu, F. and Li, X., 2014. Transfer learning for visual categorization: A survey. IEEE transactions on neural networks and learning systems, 26(5), pp.1019-1034.

[65]. Yang, B., Cao, J., Ni, R. and Zhang, Y., 2017. Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. IEEE access, 6, pp.4630-4640.

[66]. Shaha, M. and Pawar, M., 2018, March. Transfer learning for image classification. In 2018 second international conference on electronics, communication and aerospace technology (ICECA) (pp. 656-660). IEEE.

[67]. Selimović, A., Meden, B., Peer, P. and Hladnik, A., 2018, July. Analysis of content-aware image compression with VGG16. In 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI) (pp. 1-7). IEEE.

[68]. Gopalakrishnan, K., Khaitan, S.K., Choudhary, A. and Agrawal, A., 2017. Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection. Construction and building materials, 157, pp.322-330.