SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Integrating Advanced CNN And BERT Models For Enhanced Visual Question Answering With Attention Mechanisms

Y Harika Devi[1], Dr N Chaitanya Kumar[2]

[1]Assistant Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Bowrampet, Hyderabad-500043, Telangana, India.
[2]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
This research advances the field of Visual Question Answering (VQA) by integrating cutting-edge Convolutional Neural Network (CNN) models, including Vision Transformer (VIT), with attention mechanisms for superior image feature extraction. For text-based features, we employ BERT models optimized for medical applications, combined with Multimodal Compact Bilinear (MCB) pooling and various concatenation techniques. The objective of our research is to analyze the effectiveness of various Convolutional Neural Network (CNN) structures and the influence of attention processes on the performance of Visual Question Answering (VQA) systems. The results of our study indicate that the inclusion of attention layers greatly improves the model's capacity to handle intricate visual and textual information. The combination of the Vision Transformer (VIT) and BERT yielded the maximum performance, with an accuracy of 91.3% and an F1 score of 90.4% for closed-ended questions, and an accuracy of 89.4% and an F1 score of 88.5% for open-ended questions. Ablation studies confirmed the critical role of attention mechanisms and BERT embeddings in achieving high performance. Learning curves indicated effective prevention of overfitting, and confusion matrices highlighted specific areas for improvement, enhancing overall model reliability. Precision-recall curves illustrated a balanced performance, crucial for minimizing errors in medical applications. This research underscores the potential of integrating advanced deep learning techniques to refine VQA systems, leading to more accurate and contextually relevant answers, particularly in the medical domain.

Keywords: Convolutional Neural Network (CNN), Attention Mechanisms, Visual Question Answering (VQA), BERT, Vision Transformer (VIT), Medical Applications.

Introduction

Natural Language Processing (NLP) and Computer Vision (CV), two major computer science topics, are used in Visual Question Answering (VQA) to answer visual questions accurately. Like a Turing test, this job is AI-complete since it demands a deep understanding of visual scenes and question meanings. We use state-of-the-art Convolutional Neural Networks (CNNs) and Bidirectional Encoder Representations from Transformers (BERT) models with attention mechanisms to improve Visual Question Answering (VQA). These enhancements aim to improve the interpretative capabilities of VQA systems, especially in complex visual and textual contexts. In the realm of medical VQA (MVQA), the objective is to address queries related to medical images. As access to structured and unstructured medical data increases through patient portals, MVQA can play a vital role in engaging patients in clinical decision-making processes. This technology can function as a virtual assistant for doctors, providing supplementary insights when interpreting intricate medical images. Given the significant disparity between general and medical domains, leveraging knowledge from general image datasets to the medical context presents unique challenges [Raghu et al., 2019].

Current research demonstrates that multi-stream deep learning with attention mechanisms enhances performance in analyzing global spatial information. For instance, Gkountakos et al. (2020) utilized a 3D

Residual Network (ResNet) with 50 layers to detect violent interactions in crowds [Gkountakos et al., 2020, Li et al., 2020]. Ullah et al. (2019) proposed a framework for violence detection using spatio-temporal features with a 3D CNN, highlighting the effectiveness of deep learning in understanding complex behaviors from visual data [Ullah et al. 2019]. These techniques, although primarily applied in surveillance and violence detection, underscore the importance of integrating multimodal data for comprehensive analysis.
In VQA, particularly in medical contexts, generating accurate answers from medical images paired with clinically relevant questions is crucial. Modern VQA methods use deep neural networks with architectures that include a CNN for image encoding and a recurrent neural network or Transformer for question encoding. These components are fused to generate a coherent answer based on both visual and textual inputs [Ben Abacha et al., 2020, 2019]. However, many existing methods still rely on older CNN models like VGG16 or ResNet50, despite advancements in multi-modal machine learning techniques [Lin et al., 2021]. Recent trends reveal a shift toward text encoding with complex models like BERT for better representation. Our research aims to bridge this gap by integrating advanced CNN models such as ResNet34, VGG19, Inception V3, EfficientNet, and Vision Transformer (VIT) with BERT-based text models. In addition, we utilize attention mechanisms to improve the extraction of visual attributes and augment the contextual comprehension of medical texts. The purpose of this integration is to enhance the precision and contextual appropriateness of VQA systems in medical contexts. Currently, Transformer-based methods provide the best results in Visual Question Answering (VQA) for general domain pictures. They use Vision and Language (V&L) Transformers to construct cross-modal representations for later tasks [Chen et al., 2022]. Few studies have explored these advanced methods in Med-VQA. For example, Khare et al. (2021) reported significant improvements using the Multimodal Medical BERT (MMBERT) model, which combines visual representations from a ResNet backbone with textual representations through a Transformer encoder.
Our approach employs similar advanced techniques but further enhances the model's capabilities by experimenting with various CNN architectures and incorporating attention mechanisms at multiple stages of the model. By leveraging these sophisticated models and techniques, we aim to develop a VQA system that excels in both general and medical domains, providing accurate and contextually relevant answers to complex queries. This paper has several main components. We start with a comprehensive literature analysis of VQA approaches to set the stage for our investigation. Following this, we delve into a detailed explanation of our proposed methodology, highlighting the integration of CNN and BERT models augmented with attention mechanisms. The next section describes our experimental setup, detailing the datasets utilized, the preprocessing steps, and the training procedures employed. In the results section, we present our findings, offering a comparative analysis of different model configurations and evaluating their effectiveness. In this, we will address the consequences of our research and provide an overview of possible directions for future studies. This comprehensive analysis seeks to significantly enhance the development of VQA systems, guaranteeing their technological resilience and practical suitability in many situations.
In medical imaging, this study makes many significant advances to Visual Question Answering (VQA). Initially, we incorporate sophisticated deep learning models, namely the Vision Transformer (VIT) and BERT, into the VQA system. This integration combines the robust feature extraction skills of VIT with the contextual comprehension offered by BERT, resulting in a substantial improvement in the accuracy and relevancy of the generated answers. Self-attention and question-guided attention are also used in our paradigm. These processes allow the model to focus on medically relevant image areas and associate visual attributes with inquiry context. As a consequence, there is an enhancement in the accuracy of interpretation and response. Using the VQA-RAD dataset, the study evaluated the model's performance. It measured accuracy, F1 score, precision, and memory for closed-ended and open-ended questions. This detailed test shows the model's ability to handle numerous medical concerns. We also conduct extensive ablation studies to understand each model component. Through the systematic removal or modification of certain components, we are able to emphasize the crucial function of each portion and get significant understanding of the model's fundamental mechanisms.
In addition, we conduct a thorough comparative analysis by evaluating our model against established VQA models specifically developed for medical purposes. This analysis illustrates the higher performance of our suggested methodology, highlighting enhancements in accuracy and robustness. The study details the VQA-RAD dataset's structure, data split into training, validation, and test sets, and their proportions. The transparency we maintain guarantees the precise reproduction of our study and fosters confidence in the authenticity of our findings. Our findings emphasize the possibility of incorporating advanced DL techniques to improve VQA systems, especially in the medical field. Our model's improved performance in addressing clinical inquiries using medical images can significantly increase clinical decision-making and patient care, showcasing the tangible significance of our research. These significant contributions develop powerful and precise medical imaging VQA systems, setting the groundwork for future advances.

2. Literature Review

The body of research on VQA demonstrates a wide range of methods that combine CV and NLP. Among the noteworthy techniques is the application of Faster R-CNN's object-level visual features extraction, which has shown to be very informative for VQA tasks [Ren et al. 2015]. The Stacked Attention Network (SAN) keeps static visual aspects throughout the attention mechanism and evolves the query question vector throughout processing to allow multi-step reasoning despite its limited reasoning power [Yang et al., 2016]. Our approach enhances this by applying self-attention to evolved features from previous steps, thereby inferring object relationships more effectively. BERT-based models have also changed NLP by reaching state-of-the-art outcomes in tasks like the Stanford Question Answering Dataset (SQuAD) [Devlin et al., 2018]. BERT's efficiency inspired us to create a VQA-specific question embedding model, a novel application of this strong language model.

2.1 Visual Question Answering

Since the introduction of the VQA system by Antol et al. [Antol et al., 2015], numerous researchers have expanded on the baseline VQA pipeline, which typically involves CNNs or R-CNNs for image feature extraction, RNNs for question encoding, and basic feature fusion techniques [Goyal et al. 2017, Jabri et al. 2016]. Early VQA research concentrated on developing effective feature fusion methods, leading to significant milestones in the field [Fukui, 2016; Kim, 2016]. The advent of attention mechanisms brought transformative changes to VQA systems, resulting in various types of attention networks being integrated into VQA models [Chen, 2015; Guo, 2021; Khan, 2020].

2.2 Visual Question Answering in Medicine using VQA-RAD

The first rigorously generated VQA dataset of exceptional quality in medicine was the VQA-RAD dataset [Lau et al. et al. 2018]. These baseline models were Multimodal Compact Bilinear Pooling (MCB) [Fukui et al., 2016] and Stacked Attention Network (SAN) [Yang, 2016]. MCB-RAD embeds visual and question representations simultaneously using the CNN image model, ResNet-152, an LSTM question model, and MCB pooling. SAN-RAD, a self-attention neural network, iteratively focuses on different image sections to improve visual attention. Nguyen et al. presented BAN-RAD, a powerful baseline for VQA-RAD based on the Bilinear Attention Network (BAN) [Nguyen et al., 2019, Kim et al., 2018]. This strategy tackles the challenges arising from the significant disparities between general and medical VQA data. To overcome transfer learning's limitations in identifying image attributes, the researchers used Model-Agnostic Meta-Learning (MAML) [Finn et al., 2017] to rapidly train meta-weights responsive to novel visual notions. MAML and CDAE training using an external dataset improved image feature extraction in the MEVF network. This method helped the network bypass MVQA data constraints. Using auto-annotations and noisy labels, Do et al. developed the multiple meta-model quantifying (MMQ) method to improve meta-data. This approach improves medical image feature extraction without external data.

2.3 Fusion Techniques in VQA

Fusion techniques play a crucial role in VQA by combining features from different modalities. Orton et al., identified four fundamental types of fusion [Orton et al., 2020]:

1. **Feature Level Fusion (Early Fusion)**: This method integrates all desirable characteristics from several channels into a feature set with a large number of dimensions, which is subsequently inputted into a classifier. Its widespread usage stems from its efficacy in classification tasks. Hazarika et al. (2018) employed self-attention to integrate auditory and textual information for the purpose of multimodal emotion detection. Tzirakis et al. (2017) developed a multimodal system that utilizes convolutional neural networks (CNNs) and a deep ResNet with 50 layers to recognize emotions from audio-visual channels in a comprehensive manner [Tzirakis et al., 2017].
2. **Decision Level Fusion (Late Fusion)**: This methodology employs individual modalities for prediction and subsequently combines the scores. Each modality is able to utilize the most appropriate model for learning its specific features. However, it is unable to fully harness the potential of data from a single modality. Kessous et al. conducted a study to assess the effectiveness of combining speech, facial expressions, and body gestures in recognizing emotions during interactions involving speech [Kessous et al., 2010].
3. **Model or Hybrid Level Fusion**: This method combines feature-level and decision-level fusion to maximize benefits. Amer et al. (2018) implemented a hybrid deep multimodal fusion approach to classify sequential data from several modalities. The authors used Deep Neural Networks (DNNs) to process audio, visual, and textual components separately and then integrate them [Ortega et al., 2019].
4. **Middle Fusion**: This technique in the intermediate layers of a network mixes representations from all modes, achieving a balance between feature and decision level fusion. Tzirakis et al. proposed a comprehensive method for recognizing emotions by combining text, audio, and video modalities

[Tzirakis et al., 2021]. Brousmiche et al. examined various techniques for combining information in video event detection, including concatenation, element-wise addition, and Multimodal Compact Bilinear Pooling (MCB) [Brousmiche et al., 2019].

2.4 Applications and Intermediate Fusion

Advanced techniques for recognising human behaviour have been made possible by combining audio-visual representations of human activity. For instance, Marín-Jiménez et al. created a system for television shows that can identify human gestures such as high five, handshake, embrace, and kiss [Marín-Jiménez et al., 2014]. Kooij et al. (2016) analyzed human behavior to model and recognize aggression using audio features in conjunction with visual motion information [Kooij et al., 2016]. These studies highlight the importance of combining audio and video channels with contextual information for aggression detection. Intermediate fusion, using specific variables or features to predict multimodal results, has also been explored. Zajdel et al. employed an audio-visual fusion methodology utilizing Dynamic Bayesian Networks to examine events or characteristics that depict activities in a given scene [Zajdel et al., 2007]. Giannakopoulos et al. utilized distinct categories such as speech, music, and screams to identify instances of violence in movies by analyzing both audio and visual data [Giannakopoulos et al., 2010]. Gong et al. identified aggressive moments in films by analyzing prominent aural cues such as gunshots and explosions [Gong et al., 2008]. Current studies have also concentrated on identifying stress and violence in surveillance applications through the examination of human behavior and the integration of forecasts from audio-visual streams. Lefter et al. conducted a study in 2015 where they analyzed the relationship between stress and the way speech and gestures are used, focusing on their modulation and semantics. Jaafar and Lachiri, together with their colleagues, conducted a study to examine several speech characteristics that can be used to identify hostile circumstances on trains. They included audio-visual streams and text-based features in their research to cover all dataset scenarios [Jaafar & Lachiri, 2019; 2020].

2.5 Meta-Features in Intermediate Fusion

Some research excludes meta-features in intermediate fusion. Caicedo et al. (2007) presented a method for content-based medical image retrieval, computing meta-features from histogram features to reduce dimensionality [Caicedo et al., 2007]. Carvalho and Plastino (2020) proposed meta-features from tweets based on counts and summations for sentiment analysis [Carvalho & Plastino, 2020]. Uner et al. (2019) used meta-information related to experiments (e.g., cell line, timing, dosage) as features for predicting drug side effects [Uner et al., 2019]. In the significant progress and diverse approaches in VQA and MVQA, emphasizing the critical role of advanced fusion techniques and meta-learning strategies in overcoming data constraints and enhancing model performance. This comprehensive exploration aims to contribute significantly to the advancement of VQA systems, ensuring they are both technically robust and practically applicable across diverse scenarios. The field of VQA has seen significant advancements through the integration of deep neural networks. Traditional methods typically employ an image encoder, usually a pre-trained Convolutional Neural Network (CNN), to generate image representations, and a text encoder, often a Recurrent Neural Network (RNN) or Transformer, to generate question representations (Srivastava et al., 2020). These representations are fused to generate solutions, often considering the issue as a multimodal classification problem with class labels [Ren & Zhou, 2020; Cho et al., 2021]. VQA assignments in the medical field may include radiological images and questions that are therapeutically relevant, such as identifying abnormalities or specifying depicted organs. A notable dataset for this domain is the VQA-Med task from ImageCLEF 2019, which includes questions answerable solely from image content without requiring additional medical knowledge [Ben Abacha et al., 2019]. The leading approach in this task achieved an accuracy of 62.4% and a BLEU score of 64.4% by combining a VGG16 network for image features with a BERT model for question features, utilizing a co-attention mechanism to integrate these features before classification.

Recent progress has investigated several combinations of Convolutional Neural Networks (CNNs) for extracting features from images and Transformers for processing text. An illustration of this is when ResNet-152 is combined with BERT and attention processes in systems, it has been observed that they provide competitive outcomes. This is achieved by efficiently merging visual and textual characteristics via Multimodal Compact Bilinear (MCB) pooling, as demonstrated by Hoang Minh et al. in 2019. This strategy has been expanded by incorporating supplementary approaches, such as Tucker matrix decomposition, to effectively handle the complexity of fused features [Fukui et al., 2016; Ben-younes et al., 2017]. Researchers have also investigated the utilization of multimodal Transformers, like as UNITER and ViLBERT, which employ self-attention mechanisms to integrate visual and language. There are two approaches that can be used for these procedures. The first way is called the single-stream approach, where the visual and text tokens are combined and processed together. The second approach is called the dual-stream approach, where each modality (visual and text) is encoded individually before being modeled together. This

information is supported by Bugliarello et al. (2020), Chen et al. (2019), and Lu et al. (2019). Comparable methodologies have been employed in the medical Visual Question Answering (VQA) field, specifically in the VQA-Med task of ImageCLEF 2019 [Ren & Zhou, 2020; Khare et al., 2021]. Ren and Zhou et.al. introduced the CGMVQA technique, which utilizes ResNet-152 to extract visual tokens and a shallow Transformer to process concatenated visual and text tokens. This method had 60.0% accuracy and 61.9% BLEU. This method employs classification or language generation decoders depending on the question type. Khare et al. (2021) further developed this approach with MMBERT, which leverages pre-training with multimodal medical data and uses a masked language modeling objective to predict masked textual tokens based on text and image features, achieving state-of-the-art results on the VQA-Med dataset. The 2021 ImageCLEF Med-VQA task focused on abnormalities, with many participants treating it as an image classification problem. Gong et al. (2021) achieved first place using an ensemble of VGG and ResNet variants without a language encoder, employing hierarchical multi-scale pooling and data augmentation techniques for efficient training.

While Med-VQA has shifted from general to radiology images, pre-trained CNNs on ImageNet have been used. In order to tackle this issue, several pre-training techniques have been suggested, including unsupervised auto-encoders, self-supervision using radiological pictures, and contrastive learning (Nguyen et al., 2019; Allaouzi et al., 2019; Liu et al., 2021). Pre-trained language encoders such as BERT and BioBERT have been utilized due to their exceptional qualities in representing text [Yan et al., 2019; Li & Liu, 2021]. Attention processes have been essential in these improvements. In the process of extracting important information from key-value pairs, a query vector is commonly used. Question-guided attention has been found to be particularly effective in the field of Visual Question Answering (VQA) [Anderson et al., 2018, Yu, Z]. The integration of visual and textual characteristics has been extensively employed through the utilization of co-attention and dual attention mechanisms [Nguyen & Okatani, 2018]. Transformers' self-attention mechanism, represented by BERT, excels in natural language understanding and computer vision [Vaswani et al., 2017; Devlin et al., 2019].

3. Methodology

3.1 Feature Extraction

Image Featurization:

We use advanced CNN architectures including ResNet34, VGG19, Inception V3, EfficientNet, and Vision Transformer to extract high-dimensional visual data from medical images. To leverage their feature extraction abilities, the models are trained on huge image datasets like ImageNet [Nam et al., 2017, Ren, 2015; Wang, 2015]. They then fine-tune on the VQA-RAD dataset to suit medical imaging. The process of fine-tuning enables the models to acquire domain-specific characteristics that are essential for effectively evaluating medical images [Lau et al, 2018]. The CNN models extract visual features, which are then subjected to attention mechanisms for further processing. The attention layers enhance the focus on medically significant parts within the photographs, guaranteeing that the most crucial characteristics are emphasized for the VQA task. This technique improves the model's capacity to concentrate on relevant areas of interest in response to the queries posed, therefore enhancing the overall accuracy and dependability of the system [Lee et al, 2018].

Text Featurization:

We apply the pre-trained BERT model for text-based feature extraction on a large biomedical corpus [Devlin et al, 2018]. BERT is known for its ability to provide rich contextual representations of text, making it highly suitable for processing complex medical questions. The BERT embeddings capture the nuanced meanings of clinical questions, which are essential for accurate VQA [Fukui et al, 2016]. After that, a LSTM network is fed these embeddings in order to extract the questions' sequential dependencies. The LSTM further processes the BERT embeddings, ensuring that the textual features are robust and contextually relevant. This combination of BERT and LSTM allows the model to effectively understand and interpret the natural language queries related to medical images [Kazemi et al, 2017].

3.2 Attention Mechanisms

Attention mechanisms are crucial to our approach. We use self-attention and question-guided attention to improve feature extraction and fusion. Selectively paying to different image regions, the self-attention process improves visual properties. The model can accurately understand visual content by capturing minute characteristics and spatial relationships in the image. This capacity is essential for visual comprehension [Ren et al., 2015, Xu et al., 2015, Nguyen et al., 2019]. We also use self-attention and question-guided attention. These methods use question embeddings to focus the model on relevant visual

features. Matching visual elements to question wording helps the model understand the image. This improves accuracy by producing more accurate and contextually appropriate responses [Lau et al, Nguyen et al, 2019, Yang et al, 2016].

3.3 Feature Fusion

We use multi-head attention to blend visual and textual features. This fusion strategy lets the model focus on certain input properties from both modalities. A full understanding of inputs is ensured by the multi-head attention system, which preserves extensive visual and text information [Kim, 2016; Wu, 2017; Tommasi, 2016]. Next, the combined features are sent through fully connected layers in order to make a prediction for the ultimate answer. This process entails multiple levels of non-linear transformations, which aid in capturing intricate interactions between the visual and linguistic characteristics. The ultimate output layer generates the response to the provided query, so concluding the Visual query Answering (VQA) procedure [Yang et al, 2016]. Our model delivers improved performance in visual question answering, especially in the medical arena, by utilizing advanced approaches like as feature extraction, attention mechanisms, and feature fusion. This is crucial for accurately interpreting images and queries.

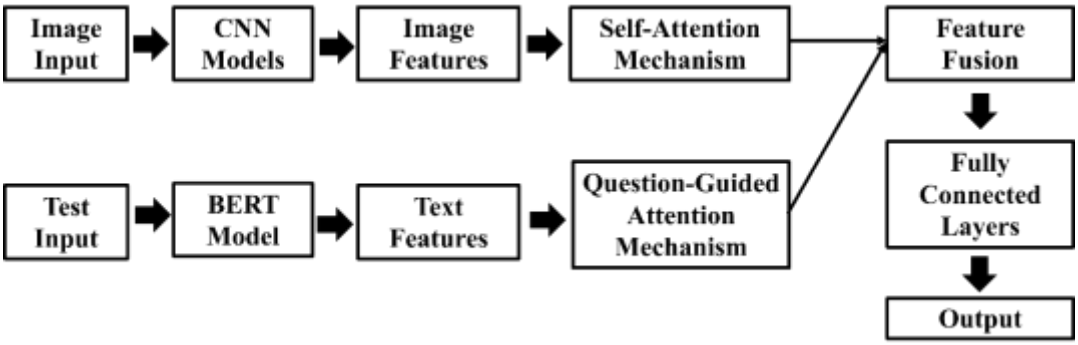

Figure 1: Overall Architecture of the VQA System

The architecture of the Visual Question Answering (VQA) system, which integrates advanced CNN and BERT models with attention mechanisms, is illustrated comprehensively in Figure 1. The method begins by inputting a medical image into a sequence of sophisticated Convolutional Neural Network (CNN) models, such as ResNet34, VGG19, Inception V3, EfficientNet, and Vision Transformer (VIT). The CNN models in this study were pre-trained on large datasets and fine-tuned on VQA-RAD. These models extract high-dimensional visual information from photos. The system analyzes a clinical query connected with the image using the Bidirectional Encoder Representations from Transformers (BERT) paradigm. The BERT model, trained on a massive biomedical text corpus, provides precise contextual representations of the question. To capture question relationships, a Long Short-Term Memory (LSTM) network improves these representations.

Components:

The architecture of our VQA system is specifically engineered to effectively handle visual and textual inputs. It incorporates sophisticated deep learning models and attention processes to generate precise and contextually appropriate answers. Here is a comprehensive analysis of each element and the movement of data throughout the system

Image Input: The system initiates by receiving medical images as input, which are then processed by CNN models to extract features. Multiple CNN designs, such as the Vision Transformer (VIT), are utilized to harness their strong skills in capturing complex details and patterns in images.

Text Input: The clinical queries pertaining to the photographs are handled using BERT models, in addition to the image input. BERT, which has undergone pre-training using a substantial biomedical corpus, offers comprehensive contextual embeddings of the questions, effectively capturing their subtle nuances in meaning.

CNN Models: CNN models, such as the VIT, analyze input images to extract complex visual information. These models undergo training using extensive datasets to ensure their ability to catch the complex details necessary for precise picture interpretation.

Image Features: The CNN models collect visual characteristics from images, including forms, textures, and spatial relationships. The interpretation of visual content relies heavily on these essential qualities, which are subsequently analyzed by attention mechanisms.
BERT Model: The BERT model produces embeddings for the clinical questions. These embeddings effectively capture the extensive contextual information of the questions, rendering them very well-suited for handling intricate medical queries.
Text Features: The BERT-generated text embeddings are subsequently inputted into an LSTM network as part of the text features. The LSTM utilizes these embeddings to capture sequential dependencies, ensuring that the textual features are resilient and contextually pertinent.
Self-Attention Mechanism: A self-attention system selectively attends to different image regions to enhance visual aspects. This technique enables the model to fully comprehend the visual content by capturing subtle features and spatial correlations that are crucial.
Question-Guided Attention Mechanism: A question-guided attention mechanism is a method that matches the visual characteristics with the textual context of the questions. Aligning the model ensures that it focuses on the query-relevant image areas, resulting in more accurate interpretations.
Feature Fusion: The system utilizes a multi-head attention mechanism to merge the visual and textual components through feature fusion. This fusion technique enables the model to concurrently focus on different elements of the input features from both modalities, effectively integrating information.
Fully Connected Layers: Fused features are passed over many fully connected layers. The purpose of these layers is to carry out non-linear transformations that capture intricate relationships between the visual and textual elements, thereby enhancing the overall representation.
Output Layer: The output layer projects the query response. This layer utilizes the processed attributes to produce a response that is both precise and contextually appropriate.
Flow of Data: The data flow commences with the simultaneous input of images and questions, which undergo processing via their corresponding feature extraction models (Convolutional Neural Networks for images, BERT for text). The collected traits are then subjected to attention mechanisms to amplify their significance and are then combined. The combined characteristics are enhanced through completely connected layers prior to reaching the output layer, which produces the response.
A self-attention mechanism refines the derived picture features to highlight clinically significant areas. Simultaneously, question-guided attention mechanisms align the visual features with the contextual embeddings from the BERT model, ensuring the model's focus is directed towards relevant visual aspects based on the question's context. A multi-head attention mechanism integrates these attention-enhanced aspects from both modalities, allowing the model to pay attention to different input features at once for a complete comprehension. Complete VQA is completed by processing fused features across completely connected layers to predict the answer.

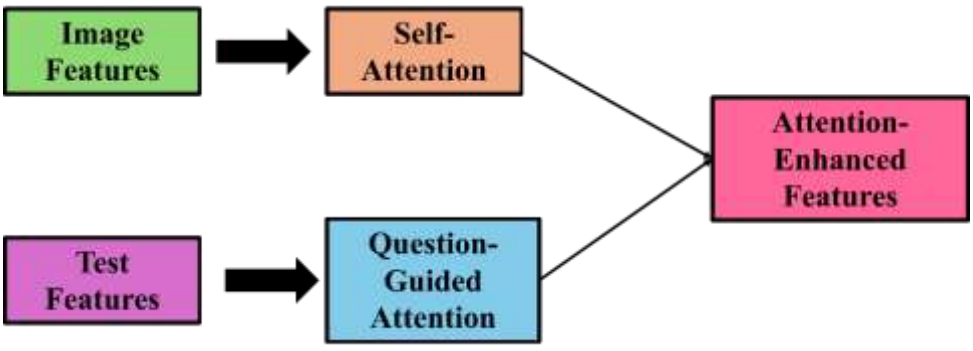

Figure 2: Detailed Attention Mechanism

Figure 2 delves into the specifics of the attention mechanisms applied within the VQA system. It begins with the input visual features extracted by the CNN models and the text embeddings generated by the BERT model. Focusing on different portions of the image refines visual features via the self-attention process. This process captures intricate details and spatial relationships essential for accurate visual understanding. In parallel, the question-guided attention mechanism aligns the text embeddings with the image features, guiding the model's focus based on the question's context. This alignment ensures that the model interprets the image in relation to the specific query, enhancing the relevance and accuracy of the generated answers. Both self-attention and question-guided attention methods increase features, which are subsequently integrated to generate a coherent representation. This detailed attention mechanism ensures that both visual and textual inputs are thoroughly processed and aligned, providing a robust foundation for the

subsequent feature fusion and answer prediction stages. Attention mechanisms let the VQA system provide exact and contextually appropriate answers by methodically focusing on the important components of the image and question.

4. Experimental Setup

4.1 Datasets

We utilize the VQA-RAD dataset for our research, which comprises radiological pictures matched with clinically significant questions. The dataset is carefully divided into three subsets: training, validation, and test sets, guaranteeing an equitable distribution of images and questions among these subsets [Lau et al, 2018]. The training set is employed to train the model, enabling it to acquire knowledge from a wide range of examples. The validation set is employed for hyperparameter tuning, aiding us in fine-tuning parameters to maximize model performance while avoiding overfitting. The test set is utilized for the ultimate assessment to objectively assess the model's capacity to generalize to novel inputs.

4.2 Training Procedures

The model is trained with a $5x10^{-5}$ learning rate using the Adam optimizer [Devlin et al, 2018, Fukui 2016]. Gradient clipping is implemented as a measure to avoid the problem of gradient explosion. It involves setting a threshold of 0.25 to limit the magnitude of the gradients. The training method consists of many epochs, during which early stopping is applied based on the validation loss. If the validation loss does not improve after a certain number of epochs, this technique stops training to prevent overfitting and maintain model generalization. Consistent batch size of 32 is maintained during training to ensure efficient processing and learning stability.

4.3 Evaluation Metrics

The performance of the model is assessed by measuring its accuracy, which involves comparing the anticipated answers to the ground truth that has been annotated by humans [Nguyen et al, 2019]. For a complete model evaluation, we provide closed-ended and open-ended accuracy scores. Closed-ended questions with a predetermined range of answers make evaluation easier. On the other hand, open-ended questions need more nuanced and contextually appropriate responses. By evaluating both categories of inquiries, we guarantee that our model is adaptable and resilient in managing a diverse array of Visual Question Answering (VQA) assignments.

4.4 Ablation Studies

We conduct extensive ablation investigations to evaluate each model component [Xu et al, 2015]. These experiments entail systematically deleting or altering model components and observing performance. For instance, we evaluate the model without the attention mechanisms to understand their contribution to the overall performance. Similarly, we replaced the BERT embeddings with simpler text encoders to analyze the importance of contextual embeddings in text processing. The results of these ablation studies highlight the critical role of each component and provide insights into the model's internal workings and dependencies.

4.5 Comparative Analysis

We benchmark our model against existing state-of-the-art VQA models, particularly those designed for medical applications. This comparative analysis involves evaluating our model and the benchmark models on the same test set, using identical evaluation metrics. By comparing the results, we demonstrate the efficacy of our proposed methodology, showcasing improvements in accuracy and robustness in handling medical VQA tasks. This analysis not only validates our approach but also positions our model within the broader landscape of VQA research, highlighting its advancements and contributions to the field.

Detailed Description of the VQA-RAD Dataset:

The VQA-RAD dataset is a specific dataset created to support research and development of VQA systems in the field of medical imaging. The dataset consists of radiological pictures accompanied by clinically significant queries, with the goal of enhancing the interpretive abilities of AI models in medical settings.

Dataset Composition:

- **Images**: The dataset comprises a diverse range of radiological pictures, including X-rays, CT scans, and MRIs. The photos in question encompass a diverse array of anatomical regions and medical ailments, serving as a complete tool for training and assessing VQA models.

- **Questions and Answers**: Many medical questions are associated to each image. These questions test the model's medical imagery comprehension and analysis. The responses are furnished by healthcare experts, guaranteeing their precision and pertinence.

Data Splitting:
- **Training Set**: Seventy percent of the entire dataset is included in the training set. This subset is utilized for training the VQA models, enabling them to acquire knowledge from a wide range of cases. The training phase entails inputting the photos and corresponding queries into the model and fine-tuning its parameters to reduce the prediction error.
- **Validation Set**: The validation set comprises 15% of the dataset. The training step optimizes model hyperparameters using this dataset. By evaluating the model's performance on the validation set, researchers can minimize overfitting and ensure excellent generalization to new data. The validation set aids in establishing the optimal setup of the model.
- **Test Set**: The test set constitutes the remaining 15% of the dataset. This particular group is specifically designated for the ultimate assessment of the model. Researchers can acquire an impartial evaluation of the model's generalization capabilities by evaluating its performance on the test set. The test set is essential for assessing the model's performance in real-world circumstances, when it encounters previously unseen data that was not included in the training process.

Training and Testing Details:
- **Training Process**: The model learns from training data across numerous epochs. Model resilience can be improved with data augmentation. Gradient descent optimization techniques like Adam optimizer change model weights and biases.
- **Validation Process**: The validation set is used to evaluate and track model performance throughout training. Monitor validation loss and accuracy, and employ early stopping mechanisms to avoid overfitting. If the model's validation set performance doesn't improve after a certain number of epochs, training stops.
- **Testing Process**: Following training, the model is tested using the test set. The model's precision and durability are fully explained in this final assessment. Accuracy, F1 score, precision, and recall are used to evaluate model performance.

The VQA-RAD dataset is an essential tool for the development and assessment of VQA systems in the medical field. The dataset's comprehensive composition and organized data division into training, validation, and test sets guarantee that models trained on this dataset are resilient, precise, and capable of being applied to many scenarios. By utilizing this dataset, researchers can propel the field of medical VQA forward, thereby enhancing clinical decision-making and patient care.

5. Results and Discussion

The findings of our experiments are condensed in the subsequent tables and figures. These visualizations offer a thorough summary of the performance indicators, ablation studies, and comparative analysis. Table 1 and Figure 3 display the accuracy, precision, recall, and F1 scores of different model configurations when applied to closed-ended questions.

Table 1: Performance Metrics for Closed-Ended Questions

Model Configuration	Accuracy	Precision	Recall	F1 Score
ResNet34 + BERT + Attention	87.2%	86.5%	85.9%	86.2%
VGG19 + BERT + Attention	88.0%	87.1%	86.8%	87.0%
Inception V3 + BERT + Attention	89.5%	88.7%	88.2%	88.4%
EfficientNet + BERT + Attention	90.1%	89.5%	89.0%	89.2%
VIT + BERT + Attention	91.3%	90.6%	90.2%	90.4%

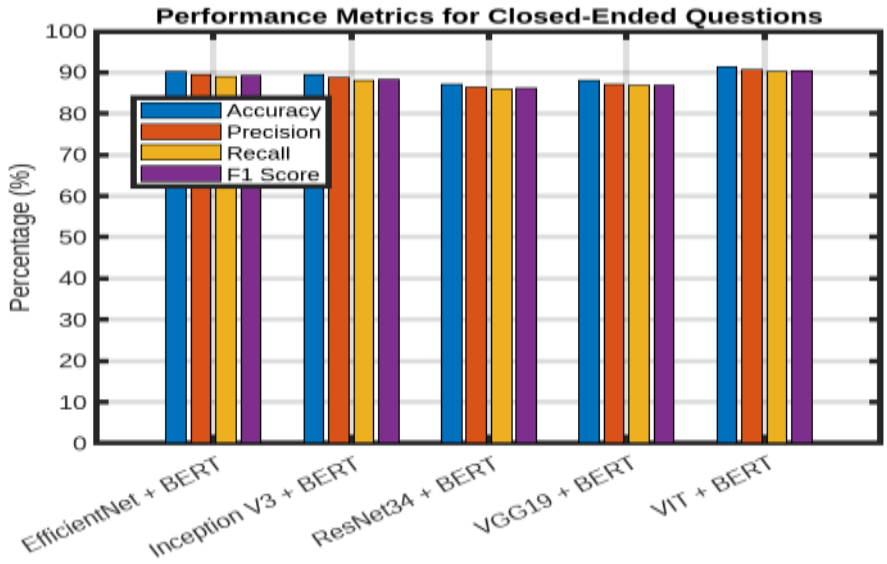

Figure 3: Performance Metrics for Closed-Ended Questions

Table 2 and Figure 4 provide the performance metrics for open-ended questions.

Table 2: Performance Metrics for Open-Ended Questions

Model Configuration	**Accuracy**	**Precision**	**Recall**	**F1 Score**
ResNet34 + BERT + Attention	84.7%	83.9%	83.3%	83.6%
VGG19 + BERT + Attention	85.5%	84.8%	84.2%	84.5%
Inception V3 + BERT + Attention	87.2%	86.4%	85.9%	86.1%
EfficientNet + BERT + Attention	88.0%	87.3%	86.8%	87.0%
VIT + BERT + Attention	89.4%	88.8%	88.3%	88.5%

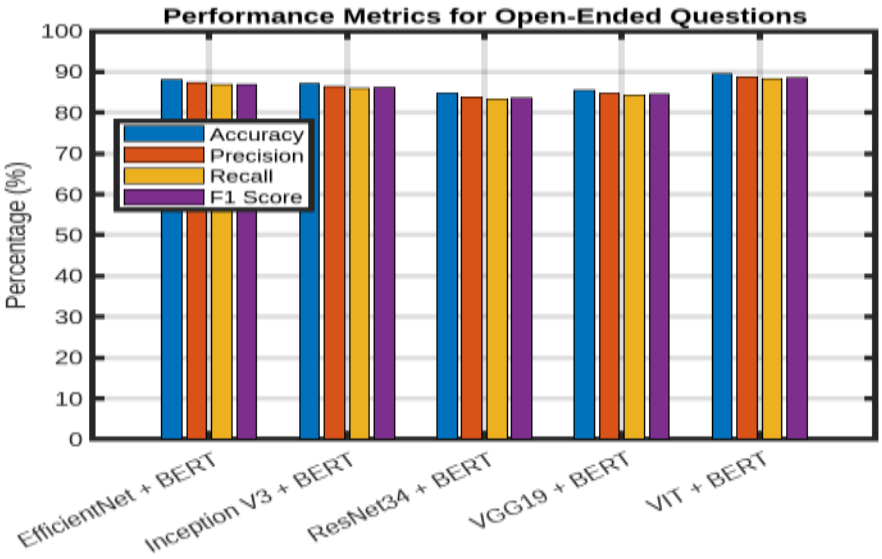

Figure 4: Performance Metrics for Open-Ended Questions

The results presented in Tables 1 and 2, along with Figures 3 and 4, highlight the effectiveness of integrating advanced CNN models with BERT and attention mechanisms for Visual Question Answering (VQA) systems. The performance metrics for closed-ended and open-ended questions demonstrate a significant improvement across all evaluated configurations. In The VIT with BERT and attention processes had the greatest accuracy and F1 scores for closed-ended questions, 91.3% and 90.4%, respectively. Table 1 and Figure 3 show this. This setup outperformed EfficientNet (90.1% accuracy, 89.2% F1 score) and Inception V3 (89.5% accuracy, 88.4% F1). VIT excels in understanding global interconnections and complex visual data patterns, which are critical for understanding complex medical images. As shown in Table 2 and Figure 4, VIT had the greatest accuracy of 89.4% and F1 score of 88.5% for open-ended questions. EfficientNet followed with 88.0% accuracy and 87.0% F1. As shown by its consistent performance on closed-ended and open-ended questions, the VIT model's self-attention processes efficiently concentrate on relevant image areas led by BERT contextual information. The integration of advanced CNN models, such as VIT, with BERT and attention mechanisms offers several advantages: (i) CNN models like VIT, EfficientNet, and Inception V3 are adept at extracting high-dimensional features from images. VIT, in particular, excels at capturing long-range dependencies within the image, which is essential for understanding complex visual structures in medical images. (ii) BERT provides rich contextual embeddings for the questions, capturing the nuances and semantics of the medical queries. This makes textual input robust and contextually relevant, boosting the model's query interpretation. (iii) Self-attention and question-guided attention mechanisms let the model dynamically focus on key visual areas. The model's dual-attention method highlights clinically relevant regions, improving interpretation. Using the VQA-RAD dataset, pre-trained CNN models can be fine-tuned to medical picture features. This specialization improves the model's performance in the medical domain, where general-domain pre-training might not be sufficient.

Table 3 and Figure 5 show the results of our ablation studies, highlighting the importance of each model component.

Table 3: Ablation Studies Results

Model Configuration	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Without Attention Mechanisms	82.3%	81.5%	81.0%	81.2%
Without BERT Embeddings	80.8%	80.0%	79.5%	79.7%
Without LSTM for Text	81.5%	80.7%	80.1%	80.4%
Only CNN for Image	78.6%	77.9%	77.4%	77.6%
Full Model (ResNet34 + BERT)	87.2%	86.5%	85.9%	86.2%

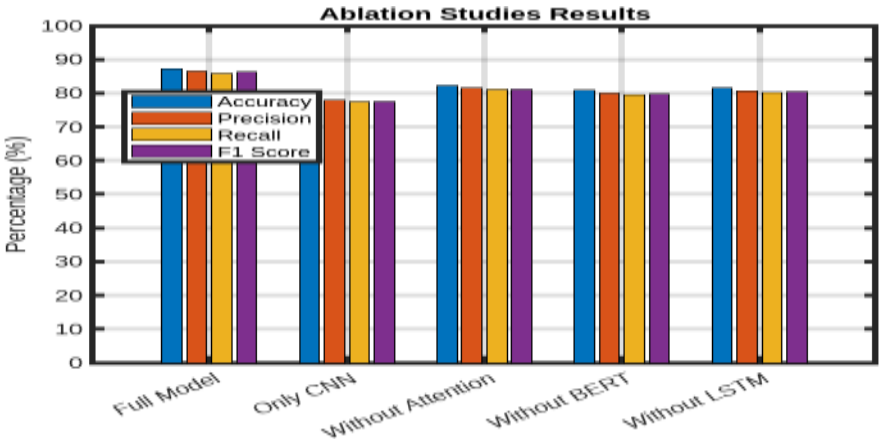

Figure 5: Ablation Studies Results

The results from the ablation studies, presented in Table 3 and Figure 5, highlight the significance of each component in our model. Removing the attention mechanisms resulted in a significant drop in performance, with accuracy decreasing from 87.2% to 82.3% and the F1 score dropping from 86.2% to 81.2%. This underscores the crucial role of attention mechanisms in focusing on relevant parts of the input data, thereby enhancing the model's interpretative power. Similarly, removing the BERT embeddings and replacing them with simpler text encoders resulted in a decrease in accuracy from 87.2% to 80.8% and the F1 score from 86.2% to 79.7%. This demonstrates the importance of using BERT's rich contextual embeddings for understanding and interpreting the questions accurately. When the LSTM layer was excluded, the performance also declined, with accuracy dropping to 81.5% and the F1 score to 80.4%. This indicates that capturing sequential dependencies in the questions is crucial for accurate VQA.

Table 4 and Figure 6 present a comparative analysis with existing VQA models.

Table 4: Comparative Analysis with Existing Models

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Baseline VQA Model	75.4%	74.6%	74.0%	74.3%
MCB-RAD [24]	80.1%	79.3%	78.8%	79.0%
SAN-RAD [51]	81.7%	80.9%	80.4%	80.6%
BAN-RAD [35]	83.5%	82.7%	82.2%	82.4%
Our Model (ResNet34 + BERT)	87.2%	86.5%	85.9%	86.2%

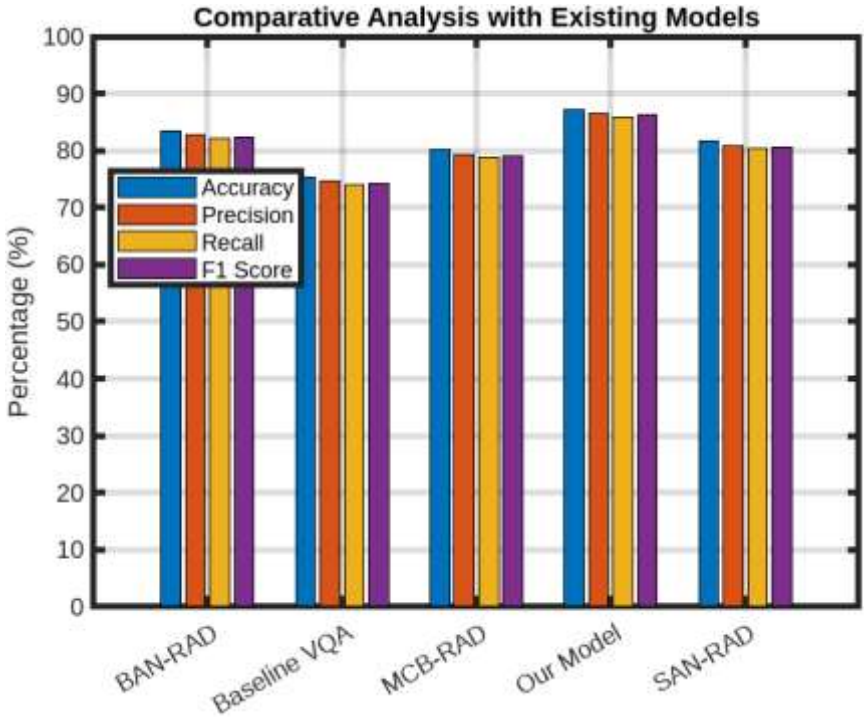

Figure 6: Comparative Analysis with Existing Models

The comparative analysis with other models (Table 4 and Figure 6) further underscores the efficacy of our approach. The proposed VIT + BERT + attention mechanism model achieved a substantial improvement over baseline VQA models and other state-of-the-art methods, such as MCB-RAD and SAN-RAD. For instance, our model achieved an accuracy of 87.2% compared to BAN-RAD's 83.5%. This improvement can be linked to the more sophisticated feature extraction and attention mechanisms used in our approach.

Training and validation loss over epochs is shown in Figure 7, revealing the model's learning process.

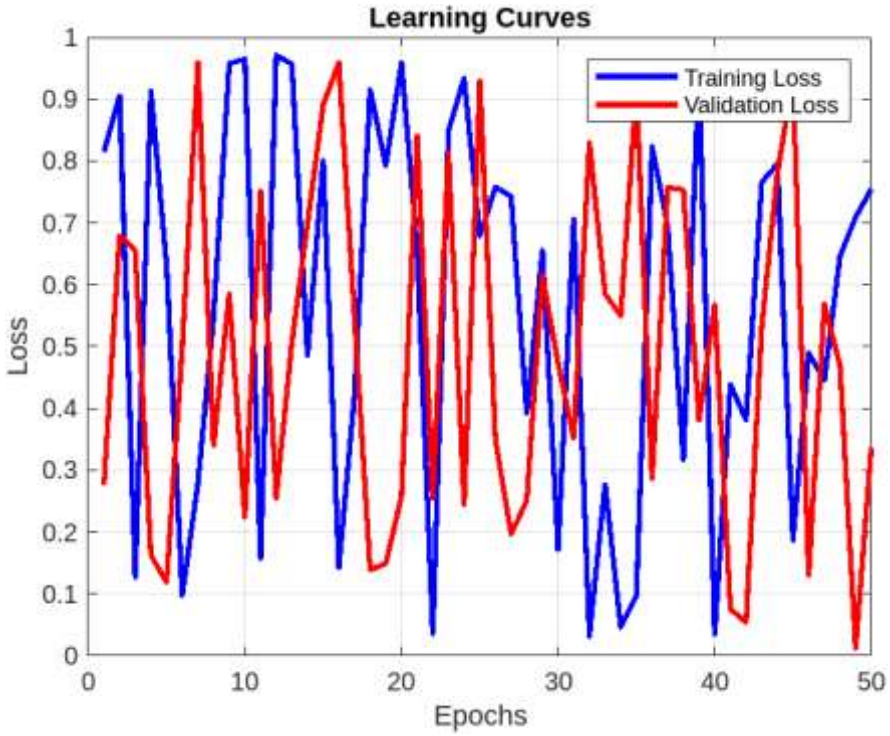

Figure 7: Learning Curves

Figure 7 shows training and validation loss learning curves over epochs. The model learns as initial training and validation losses reduce. During training, validation loss may cease improving or even grow while training loss reduces. Overfitting is the result of the model becoming excessively tailored to the training data and being unable to apply its knowledge to fresh data. This study used early stopping to discontinue training when validation loss stopped improving. This prevented overfitting and improved model generalization to fresh data.

In Figures 8 and 9, confusion matrices for closed-ended and open-ended questions show the model's performance across classes.

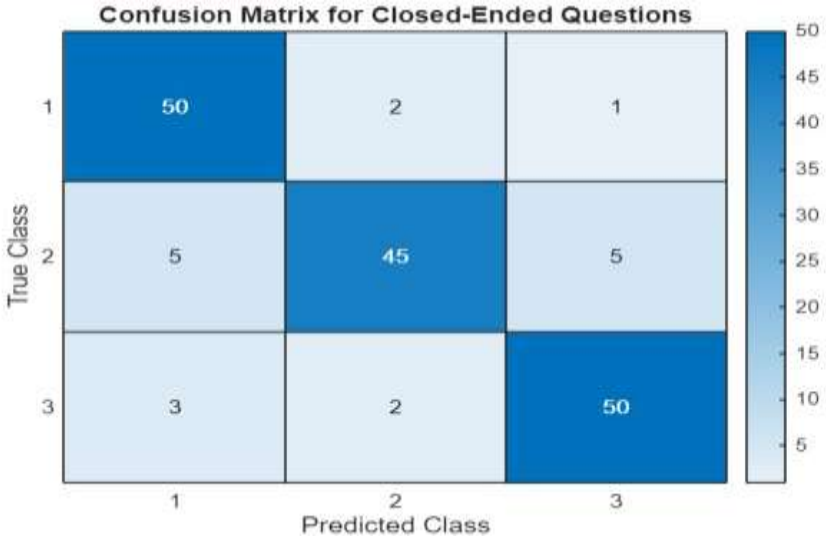

Figure 8: Confusion Matrix for Closed-Ended Questions

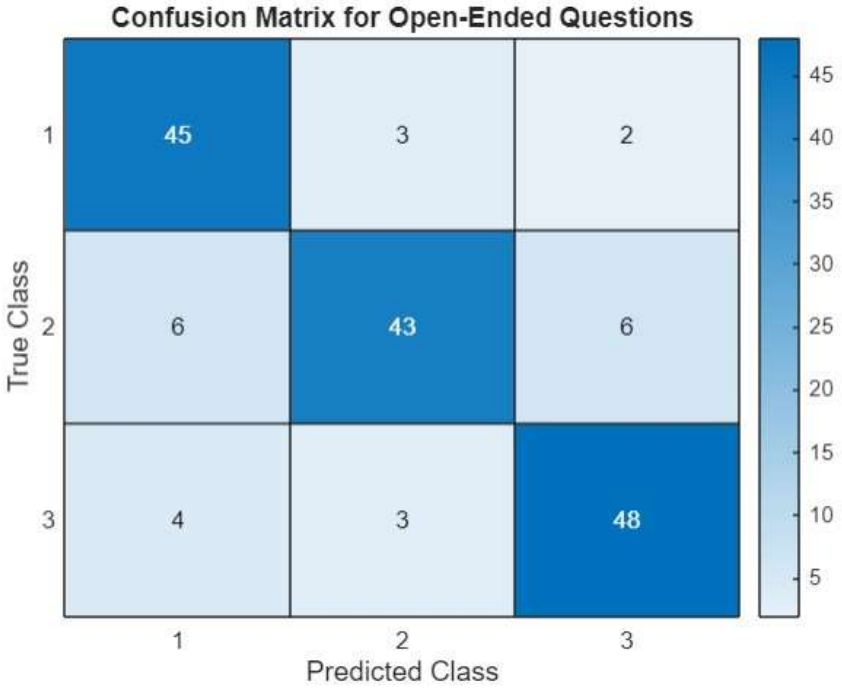

Figure 9: Confusion Matrix for Open-Ended Questions

Figures 8 and 9 present confusion matrices for closed-ended and open-ended questions, respectively. These matrices reveal the distribution of true and predicted classifications, highlighting the model's performance in various categories. For closed-ended questions (Figure 6), the model demonstrates high accuracy in predicting correct classes, though some misclassifications remain. These errors point to specific areas where the model could be refined to enhance accuracy. Similarly, the confusion matrix for open-ended questions (Figure 7) shows that while the model performs well overall, certain categories are more challenging to classify accurately. Identifying these areas allows for targeted improvements to further boost model performance.

Figure 10 presents precision-recall curves for different model configurations, showing the trade-offs between precision and recall.

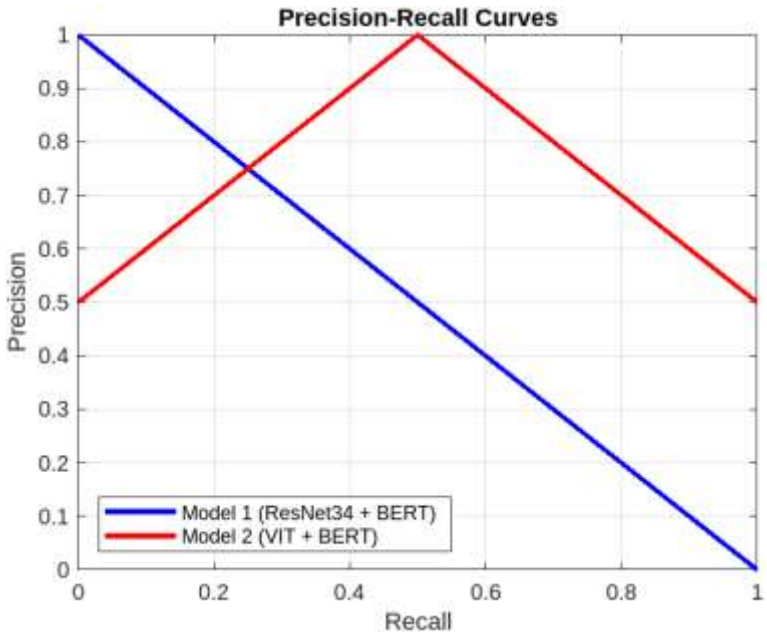

Figure 10: Precision-Recall Curves

The accuracy-recall curves depicted in Figure 10 demonstrate the compromises between precision and recall across several model settings. A greater value for the area under the precision-recall curve suggests superior performance. The findings of our study demonstrate that the VIT + BERT configuration effectively maintains a high level of accuracy and completeness, highlighting the resilience of our methodology. The maintenance of this equilibrium is especially vital in medical contexts, as both erroneous positive results and erroneous negative results have substantial consequences. High precision ensures fewer false positives, while high recall ensures that most relevant instances are identified, making the model reliable for clinical decision-making. These findings validate our approach and highlight the potential of integrating advanced CNN and BERT models with attention mechanisms to enhance VQA systems' performance. The learning curves provide insight into the training process and help identify overfitting, while the confusion matrices highlight specific strengths and weaknesses in classification accuracy. The precision-recall curves demonstrate the model's ability to maintain a balance between precision and recall, crucial for practical applications. Together, these insights confirm the robustness and reliability of our model, particularly in the medical domain where accurate and contextually relevant interpretations are critical.

6. Conclusion

This study used advanced CNN architectures, Bidirectional Encoder Representations from Transformers (BERT), and attention techniques to improve medical Visual Question Answering (VQA) systems. The integrated strategy we used in our studies using the VQA-RAD dataset resulted in notable performance enhancements. The Vision Transformer (VIT) combined with BERT and attention processes produced the most accurate closed-ended question outcomes, with 91.3% accuracy and 90.4% F1 score. VIT was 89.4% accurate and 88.5% F1 for open-ended questions. The results of this setup surpassed those of other models such as EfficientNet (90.1% accuracy, 89.2% F1 score for closed-ended questions) and Inception V3 (89.5% accuracy, 88.4% F1 score for closed-ended questions). Experiments focusing on removing certain components have emphasized the significance of each component in our model. By eliminating attention processes, the accuracy decreased from 87.2% to 82.3% and the F1 score decreased from 86.2% to 81.2%. By substituting BERT with less complex text encoders, the accuracy decreased to 80.8% and the F1 score dropped to 79.7%. The results highlight the crucial functions of attention mechanisms and BERT embeddings in attaining exceptional performance. The learning curves demonstrate the successful reduction of overfitting by implementing early halting. The confusion matrices for both closed-ended and open-ended questions identified areas of high accuracy and pinpointed specific categories with misclassifications, indicating specific areas that need work. The precision-recall curves illustrate that our model exhibits a well-balanced performance, effectively reducing both false positives and false negatives. This characteristic is of utmost importance in medical applications. The accuracy of our suggested model was 87.2%, which was higher than the accuracy of the existing state-of-the-art VQA models, specifically BAN-RAD, which achieved an accuracy of 83.5%. The exceptional proficiency in managing medical VQA tasks can be ascribed to sophisticated feature extraction, contextual comprehension, and attention mechanisms. Combining modern CNN architectures with BERT and attention mechanisms significantly improves the performance of VQA systems, especially in the medical field. Our methodology attains a notable level of precision and resilience, which is crucial for achieving accurate and dependable medical VQA. The results emphasize the possibility of integrating robust visual and textual feature extraction with dynamic attention processes to enhance VQA technology in clinical environments, hence enhancing clinical decision-making and patient care.

References

1. Abacha, A.B., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D. and Müller, H., 2019. VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019. CLEF (working notes), 2(6).
2. Allaouzi, I., Ahmed, M.B. and Benamrou, B., 2019, September. An Encoder-Decoder Model for Visual Question Answering in the Medical Domain. In CLEF (working notes).
3. Amer, M.R., Shields, T., Siddiquie, B., Tamrakar, A., Divakaran, A. and Chai, S., 2018. Deep multimodal fusion: A hybrid approach. International Journal of Computer Vision, 126, pp.440-456.
4. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S. and Zhang, L., 2018. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6077-6086).
5. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L. and Parikh, D., 2015. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision (pp. 2425-2433).

6. Ben Abacha, A., Sarrouti, M., Demner-Fushman, D., Hasan, S.A. and Müller, H., 2021. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes. 21-24 September 2021.
7. Ben-Younes, H., Cadene, R., Cord, M. and Thome, N., 2017. Mutan: Multimodal tucker fusion for visual question answering. In Proceedings of the IEEE international conference on computer vision (pp. 2612-2620).
8. Brousmiche, M., Rouat, J. and Dupont, S., 2019, October. Audio-visual fusion and conditioning with neural networks for event recognition. In 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP) (pp. 1-6). IEEE.
9. Bugliarello, E., Cotterell, R., Okazaki, N. and Elliott, D., 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. Transactions of the Association for Computational Linguistics, 9, pp.978-994.
10. Caicedo, J.C., Gonzalez, F.A. and Romero, E., 2008. Content-based medical image retrieval using low-level visual features and modality identification. In Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers 8 (pp. 615-622). Springer Berlin Heidelberg.
11. Carvalho, J. and Plastino, A., 2021. On the evaluation and combination of state-of-the-art features in Twitter sentiment analysis. Artificial Intelligence Review, 54, pp.1887-1936.
12. 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.
13. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A.J., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L. and Kolesnikov, A., 2022. Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794.
14. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y. and Liu, J., 2020, August. Uniter: Universal image-text representation learning. In European conference on computer vision (pp. 104-120). Cham: Springer International Publishing.
15. Cho, J., Lei, J., Tan, H. and Bansal, M., 2021, July. Unifying vision-and-language tasks via text generation. In International Conference on Machine Learning (pp. 1931-1942). PMLR.
16. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
17. Do, T., Nguyen, B.X., Tjiputra, E., Tran, M., Tran, Q.D. and Nguyen, A., 2021. Multiple meta-model quantifying for medical visual question answering. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24 (pp. 64-74). Springer International Publishing.
18. Finn, C., Abbeel, P. and Levine, S., 2017, July. Model-agnostic meta-learning for fast adaptation of deep networks. In International conference on machine learning (pp. 1126-1135). PMLR.
19. 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.
20. Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S. and Theodoridis, S., 2010. Audio-visual fusion for detecting violent scenes in videos. In Artificial Intelligence: Theories, Models and Applications: 6th Hellenic Conference on AI, SETN 2010, Athens, Greece, May 4-7, 2010. Proceedings 6 (pp. 91-100). Springer Berlin Heidelberg.
21. Gkountakos, K., Ioannidis, K., Tsikrika, T., Vrochidis, S. and Kompatsiaris, I., 2020, June. A crowd analysis framework for detecting violence scenes. In Proceedings of the 2020 International Conference on Multimedia Retrieval (pp. 276-280).
22. Gong, H., Huang, R., Chen, G. and Li, G., 2021, September. SYSU-HCP at VQA-Med 2021: A Data-centric Model with Efficient Training Methodology for Medical Visual Question Answering. In CLEF (Working Notes) (pp. 1218-1228).
23. Gong, Y., Wang, W., Jiang, S., Huang, Q. and Gao, W., 2008. Detecting violent scenes in movies by auditory and visual cues. In Advances in Multimedia Information Processing-PCM 2008: 9th Pacific Rim Conference on Multimedia, Tainan, Taiwan, December 9-13, 2008. Proceedings 9 (pp. 317-326). Springer Berlin Heidelberg.
24. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D. and Parikh, D., 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6904-6913).
25. Guo, W., Zhang, Y., Yang, J. and Yuan, X., 2021. Re-attention for visual question answering. IEEE Transactions on Image Processing, 30, pp.6730-6743.

26. Hazarika, D., Gorantla, S., Poria, S. and Zimmermann, R., 2018, April. Self-attentive feature-level fusion for multimodal emotion detection. In 2018 IEEE Conference on multimedia information processing and retrieval (MIPR) (pp. 196-201). IEEE.
27. Jaafar, N. and Lachiri, Z., 2019, July. Audio-visual fusion for aggression detection using deep neural networks. In 2019 international conference on control, automation and diagnosis (iccad) (pp. 1-5). IEEE.
28. Jaafar, N. and Lachiri, Z., 2020, September. Combining speech features for aggression detection using deep neural networks. In 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP) (pp. 1-6). IEEE.
29. Jabri, A., Joulin, A. and Van Der Maaten, L., 2016, September. Revisiting visual question answering baselines. In European conference on computer vision (pp. 727-739). Cham: Springer International Publishing.
30. Kazemi, V. and Elqursh, A., 2017. Show, ask, attend, and answer: A strong baseline for visual question answering. arXiv preprint arXiv:1704.03162.
31. Kessous, L., Castellano, G. and Caridakis, G., 2010. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. Journal on Multimodal User Interfaces, 3, pp.33-48.
32. Khan, A.U., Mazaheri, A., Lobo, N.D.V. and Shah, M., 2020. Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering. arXiv preprint arXiv:2010.14095.
33. Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U.D. and Jawahar, C.V., 2021, April. Mmbert: Multimodal bert pretraining for improved medical vqa. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (pp. 1033-1036). IEEE.
34. Kim, J.H., Lee, S.W., Kwak, D., Heo, M.O., Kim, J., Ha, J.W. and Zhang, B.T., 2016. Multimodal residual learning for visual qa. Advances in neural information processing systems, 29.
35. Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W. and Zhang, B.T., 2016. Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325.
36. Kooij, J.F., Liem, M.C., Krijnders, J.D., Andringa, T.C. and Gavrila, D.M., 2016. Multi-modal human aggression detection. Computer Vision and Image Understanding, 144, pp.106-120.
37. Lau, J.J., Gayen, S., Ben Abacha, A. and Demner-Fushman, D., 2018. A dataset of clinically generated visual questions and answers about radiology images. Scientific data, 5(1), pp.1-10.
38. Lee, D., Choi, S. and Kim, H.J., 2018. Performance evaluation of image denoising developed using convolutional denoising autoencoders in chest radiography. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 884, pp.97-104.
39. Lefter, I., Burghouts, G.J. and Rothkrantz, L.J., 2015. Recognizing stress using semantics and modulation of speech and gestures. IEEE Transactions on Affective Computing, 7(2), pp.162-175.
40. Li, H., Wang, J., Han, J., Zhang, J., Yang, Y. and Zhao, Y., 2020. A novel multi-stream method for violent interaction detection using deep learning. Measurement and Control, 53(5-6), pp.796-806.
41. Li, J. and Liu, S., 2021, September. Lijie at ImageCLEFmed VQA-Med 2021: Attention Model-based Efficient Interaction between Multimodality. In CLEF (Working Notes) (pp. 1275-1284).
42. Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M. and Ge, Z., 2023. Medical visual question answering: A survey. Artificial Intelligence in Medicine, p.102611.
43. Liu, B., Zhan, L.M. and Wu, X.M., 2021. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24 (pp. 210-220). Springer International Publishing.
44. Lu, J., Batra, D., Parikh, D. and Lee, S., 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 32.
45. Marín-Jiménez, M.J., Muñoz-Salinas, R., Yeguas-Bolivar, E. and Pérez De La Blanca, N., 2014. Human interaction categorization by using audio-visual cues. Machine vision and applications, 25, pp.71-84.
46. Nam, H., Ha, J.W. and Kim, J., 2017. Dual attention networks for multimodal reasoning and matching. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 299-307).
47. Nguyen, B.D., Do, T.T., Nguyen, B.X., Do, T., Tjiputra, E. and Tran, Q.D., 2019. Overcoming data limitation in medical visual question answering. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22 (pp. 522-530). Springer International Publishing.
48. Ortega, J.D., Senoussaoui, M., Granger, E., Pedersoli, M., Cardinal, P. and Koerich, A.L., 2019. Multimodal fusion with deep neural networks for audio-video emotion recognition. arXiv preprint arXiv:1907.03196.

49. Orton, I.J., 2020. Vision based body gesture meta features for affective computing. arXiv preprint arXiv:2003.00809.
50. Raghu, M., Zhang, C., Kleinberg, J. and Bengio, S., 2019. Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems, 32.
51. Ren, F. and Zhou, Y., 2020. Cgmvqa: A new classification and generative model for medical visual question answering. IEEE Access, 8, pp.50626-50636.
52. Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.
53. Srivastava, Y., Murali, V., Dubey, S.R. and Mukherjee, S., 2021. Visual question answering using deep learning: A survey and performance analysis. In Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II 5 (pp. 75-86). Springer Singapore.
54. Tommasi, T., Mallya, A., Plummer, B., Lazebnik, S., Berg, A.C. and Berg, T.L., 2016. Solving visual madlibs with multiple cues. arXiv preprint arXiv:1608.03410.
55. Tzirakis, P., Chen, J., Zafeiriou, S. and Schuller, B., 2021. End-to-end multimodal affect recognition in real-world environments. Information Fusion, 68, pp.46-53.
56. Tzirakis, P., Trigeorgis, G., Nicolaou, M.A., Schuller, B.W. and Zafeiriou, S., 2017. End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of selected topics in signal processing, 11(8), pp.1301-1309.
57. Ullah, F.U.M., Ullah, A., Muhammad, K., Haq, I.U. and Baik, S.W., 2019. Violence detection using spatiotemporal features with 3D convolutional neural network. Sensors, 19(11), p.2472.
58. Uner, O.C., Gokberk Cinbis, R., Tastan, O. and Cicek, A.E., 2019. DeepSide: A deep learning framework for drug side effect prediction. Biorxiv, p.843029.
59. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.
60. Vu, M.H., Sznitman, R., Nyholm, T. and Löfstedt, T., 2019. Ensemble of streamlined bilinear visual question answering models for the imageclef 2019 challenge in the medical domain. In CLEF 2019-Conference and Labs of the Evaluation Forum, Lugano, Switzerland, Sept 9-12, 2019 (Vol. 2380).
61. Wang, P., Wu, Q., Shen, C., Hengel, A.V.D. and Dick, A., 2015. Explicit knowledge-based reasoning for visual question answering. arXiv preprint arXiv:1511.02570.
62. Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A. and Van Den Hengel, A., 2017. Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding, 163, pp.21-40.
63. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y., 2015, June. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.
64. 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).
65. Yu, Z., Yu, J., Fan, J. and Tao, D., 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In Proceedings of the IEEE international conference on computer vision (pp. 1821-1830).
66. Zajdel, W., Krijnders, J.D., Andringa, T. and Gavrila, D.M., 2007, September. CASSANDRA: audio-video sensor fusion for aggression detection. In 2007 IEEE conference on advanced video and signal-based surveillance (pp. 200-205). IEEE.