The Journal of Supercomputing (2022) 78:15042–15059
https://doi.org/10.1007/s11227-022-04474-8

Hybrid deep learning model for answering visual medical questions

Karim Gasmi[1]

Accepted: 19 March 2022 / Published online: 11 April 2022

Abstract

Due to the increase in electronic documents containing medical information, the search for specific information is often complex and time-consuming. This has prompted the development of new tools designed to address this issue. Automated visual question/answer (VQA) systems are becoming more challenging to develop. These are computer programs that take images and questions as input and then combine all inputs to generate text-based answers. Due to the enormous amount of question and the limited number of specialists, many issues stay unanswered. It's possible to solve this problem by using automatic question classifiers that guide queries to experts based on their subject preferences. For these purposes, we propose a VQA approach based on a hybrid deep learning model. The model consists of three steps: (1) the classification of medical questions based on a BERT model; (2) image and text feature extraction using a deep learning model, more specifically the extraction of medical image features by a hybrid deep learning model; and (3) text feature extraction using a Bi-LSTM model. Finally, to predict the appropriate answer, our approach uses a KNN model. Additionally, this study examines the influence of the Adam, AdaGrad, Stochastic gradient descent and RMS Prop optimization techniques on the performance of the network. As a consequence of the studies, it was shown that Adam and SGD optimization algorithms consistently produced higher outcomes. Experiments using the ImageCLEF 2019 dataset revealed that the suggested method increases BLEU and WBSS values considerably.

Keywords Medical system · Visual question answering · Deep learning · Bi-LSTM · Hybrid model

✉ Karim Gasmi
kgasmi@ju.edu.sa

[1] Department of Computer Science, College of Arts and Sciences at Tabarjal, Jouf University, Jouf, Saudi Arabia

1 Introduction

The use of automated computer frameworks in the context of clinical practice is focused on two populations: specialists who utilize these frameworks to obtain a second opinion on their conclusions, and patients who have increasing access to thorough and definitive clinical information, which they see as confusing. Thus, with regard to patients the goal of the frameworks is to give them a better understanding of their ailments by offering broad clarifications on the discoveries made during their clinical trials and sweeps. This in turn helps the specialists, who do not have time to explain every piece of information in every patient's documents.

According to Pew Research Center's Internet and American Life Project[1], 77% of people surveyed said they first began searching for health-related information using a web platform. Question/answer (QA) is a field of artificial intelligence (AI) that can be applied to these types of challenges [1].

QA systems are designed to deliver exact responses to questions asked in natural languages, within the context of a larger framework. A response can be aggregated in different ways: a passage of text from a document collection [2] or the Internet [3], as well as data obtained from a database [4]or a knowledge base, can be used to generate a response [5]. Under rare circumstances, the returned responses contain multimedia [6]. Our paper presents a model for these circumstances, a visual question/answer (VQA) system for use in the medical field. The goal of the VQA is to provide textual responses to textual questions posed in the context of a particular image.

A VQA system is made up of three major tasks [7]: (1) medical question classification, (2) image and text feature extraction and (3) retrieval and extraction of the answers in textual form. In this context, a hybrid deep learning model is proposed for medical VQA purposes. First and foremost, we suggest a new classification scheme for medical question. The dataset queries are broken down into four categories: modality, plan, organ and anomaly. The BERT pre-trained model cited by Devlin [8] was employed as a classification approach. Since its beginning in 2018, BERT has had a great deal of success. As a result of this and other transformer-based models, the area of natural language processing has advanced significantly (NLP). It gives researchers the tools they need to achieve cutting-edge outcomes, such as text categorization and machine translation that is both fast and accurate. ResNet and VGG, two deep learning models, were used to construct a technique for extracting features from medical images. Each image was subjected to both of these models, with the characteristics retrieved by each model being merged. This was followed by a model called the Bi-LSTM for extracting text features in both directions. Finally, we utilized the softmax layer to classify data and find the best response to a specific query.

The key contributions of this paper are as follows:

[1] http://pewinternet.org/Reports/2013/Health-online.aspx.

- Medical question classification based on a BERT model
- A novel medical image feature extraction method based on a hybrid deep learning model
- VQA based on a hybrid of two models ("CNN + RNN")
- Investigates the effect of the Adam, AdaGrad, Stochastic gradient descent and RMS Prop optimization approaches on the network's performance.

The remainder of this work is structured as follows. Section 2 discusses several VQA-related works. Our procedure is described in Sect. 3. Section 4 describes the tests performed on the Medical ImageCLEF 2019 dataset and reports their results, especially in VQA tasks. Finally, Sect. 5 discusses our conclusions and some different viewpoints.

2 Related works

The concept of a question/answer (QA) system was introduced in the late 1970s with the Question Answering Mechanism (QUALM), which was developed by Lehnert in 1977 [9]. QA systems are an extension of document retrieval systems, which perform similar functions. This sort of technology enables a user to ask a question in natural language and get a precise response, rather than a collection of pages considered relevant, as search engines do [10–12].

In highly specialized fields, such as medicine, a more in-depth examination of the chosen documents is necessary in order to extract the relevant information. Thus, QA systems are distinguishable from other information retrieval systems by the complexity of their architecture. A question can take different forms and can exhibit varying degrees of knowledge. In [13], for instance, Monceaux and Robba leverage syntactic knowledge of the words used in the questions, while Mendes [14] relies on the transformation of the elements of the question into logical predicates.

Several other systems have been proposed to improve QA system performance, such as QALC [15], FRASQUES [16] QRISTAL [17] and WEBCOOP [18]. All of these systems focus solely on text features in general, nonspecific domains, thereby reducing their own efficacy. The medical field, like any other specialty field, is characterized by the complexity of its vocabulary and the specificity of its technical terminology [19, 20]. Consequently, access to medical knowledge requires special handling, especially given the structure of the various resources on the Internet.

To limit the user's search space [21] and help them find the right answer, researchers are moving from text-only systems to visual QA (VQA) systems, which combine text and images. VQA models take image- and text-based questions as input in order to arrive at the most pertinent answer. They combine natural language processing (NLP) with sophisticated computer vision to generate an accurate response to a specific question [22–24] and can answer questions on the visual content of a given image using a dataset of image-question-answer triplets.

The method presented in [25] was one of the first to use deep neural network representations to accomplish text-image alignment. It extracted 4096-dimensional vectors from a set of images using a pretrained AlexNet [26], 1000-dimensional vectors from words using a pretrained skip-gram model [27] and a pairwise loss with a linear projection between modalities using a pairwise loss with a linear projection.

In [28], the authors chose to train a multi-layer network on top of vector representations of both modalities in the same dimensional space. In [29], however, the authors integrated both modalities repeatedly until the final similarity processing stage. Similarly, the authors in [30] suggested rapid parameter adaptation for image-text modeling (FPAIT), a technique that controls image feature extraction layers by producing dynamic normalizing parameters from text features. To emulate the human thinking process, neural-symbolic (NS) techniques [31, 32] use executable symbolic programs.

VQA faces two challenges in the field of medicine: not only are medical texts and images distinct from each other, the resources and labelled data available in the medical field are limited compared to what is available in more generalized fields. Additionally, questions are formed from small quantities of words, which negatively influences the results. To address this issue, the authors in [33] recommended that the quality of multimodal representations be improved. To generate a domain-specific weight initialization for the Med-VQA system, they employed a visual feature extractor pre-trained on external medical datasets with an unsupervised auto-encoder (CDAE) [34] and a meta- learning method (MAML) [35].While these initial efforts advanced the study of Med-VQA, they primarily focused on enhancing the feature extraction module and did not explore the reasoning module, which is important in high-level reasoning tasks. To overcome the limited number of feature extractions, we studied image feature expansion. Unlike previous works, we propose a novel hybrid deep learning model for VQA. Our approach is based on a hybrid feature extraction model. The first features are extracted from images; this step is based on two deep learning models. Subsequent features are extracted from text using a bidirectional long short-term memory (Bi-LSTM) model. Following feature extraction, we propose an efficient method of classifying and extracting the optimal answer to the input question.

3 Proposed model for visual question answering

In this section, we describe our proposed visual question/answer (VQA) model. The overview of this method is presented in Fig. 1, highlighting the following steps: (1) question classification and image preprocessing; (2) image feature extraction; (3) text feature extraction; (4) classification.

3.1 Pre-processing

We used a mixed deep learning model to extract picture characteristics. This was done in order to reduce the risk of overfitting, and we raised the number of images per image to 10 in order to reduce the risk of overfitting by different data augmentation technique such as clipping, rotation and scaling. Preventative measures included text preparation procedures such as stemming and lemmatization to re-form verbs

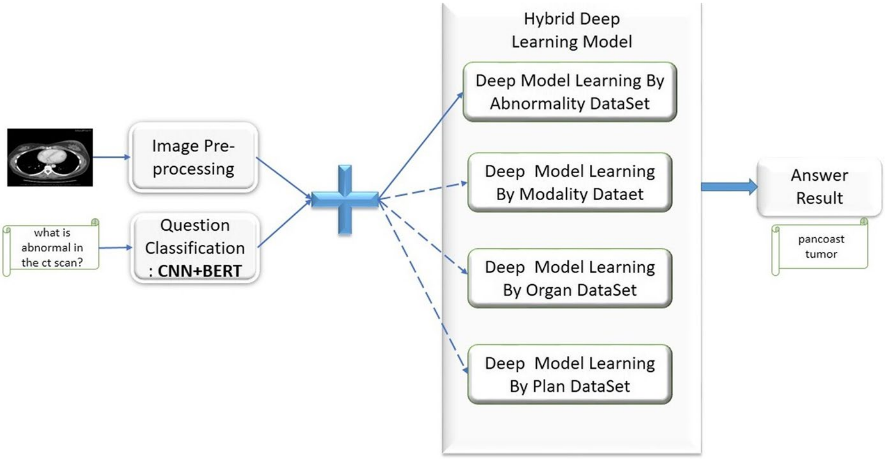
Image Pre-processing
what is abnormal in the ct scan?
Question Classification : CNN+BERT
Hybrid Deep Learning Model
Deep Model Learning By Abnormality DataSet
Deep Model Learning By Modality Dataet
Deep Model Learning By Organ DataSet
Deep Model Learning By Plan DataSet
Answer Result
pancoast tumor

Fig. 1 General proposal model architecture

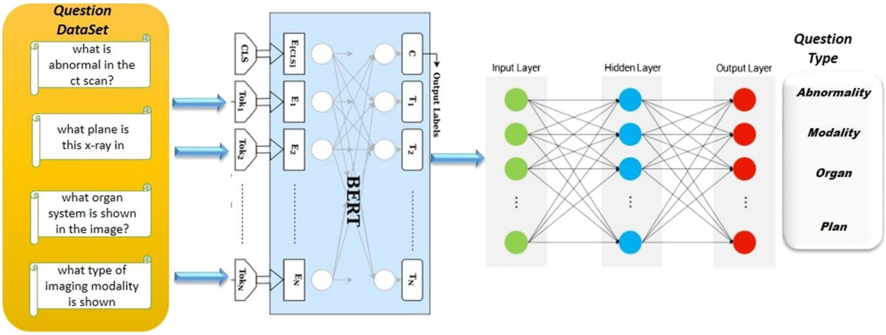

Fig. 2 Question classification

and other words in order to reduce the likelihood of their being misconstrued. All stop words and special letters were also removed from the text entirely.

After that, we attempted to convert technical medical terminology into their respective acronyms in the event that both were employed. Then, we removed low-frequency phrases based on the word frequency distribution throughout the data analysis process to guarantee that the training efficiency was maintained.

The ImageClef 2019 dataset allowed us to classify questions into four categories: plan, modality, organ and abnormality. For question feature extraction, we propose an approach based on the BERT model [8]; for classification, we combine the output of the Bidirectional Encoder Representations from Transformers (BERT) model and a fully connected layer. This approach is presented in Fig. 2.

As of late 2018, BERT is a language representation model developed by Google. That it is able to generate deep bidirectional representations from text is its biggest strength. The model may therefore learn information either from left to right or from

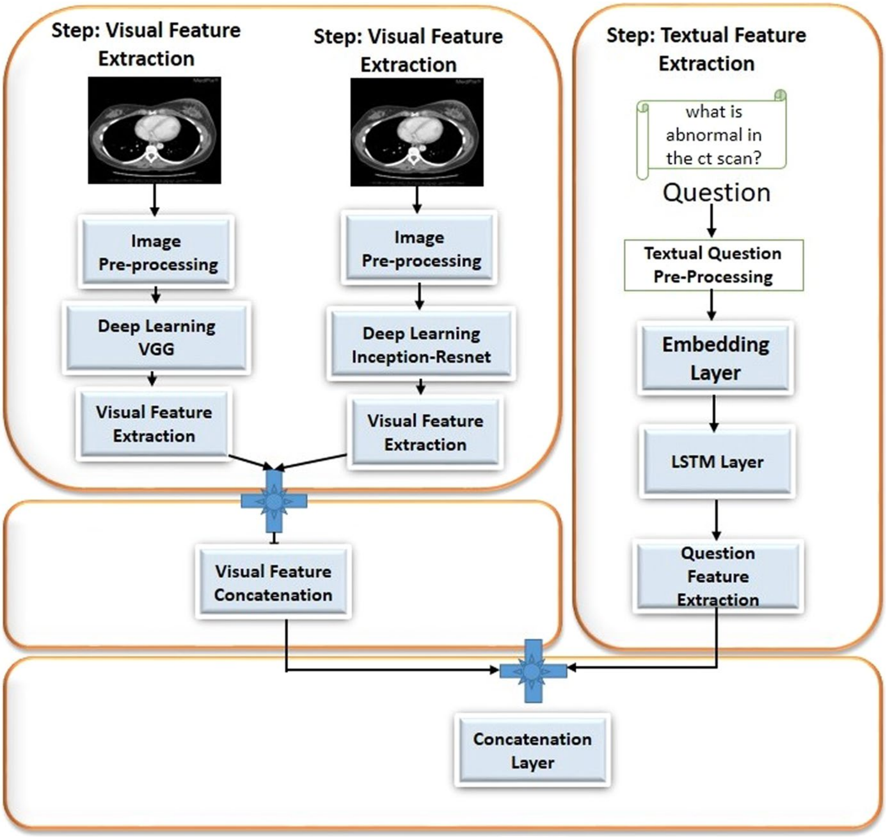
Step: Visual Feature Extraction
Step: Visual Feature Extraction
Step: Textual Feature Extraction
what is abnormal in the ct scan?
Question
Image Pre-processing
Image Pre-processing
Textual Question Pre-Processing
Deep Learning VGG
Deep Learning Inception-Resnet
Embedding Layer
Visual Feature Extraction
Visual Feature Extraction
LSTM Layer
Question Feature Extraction
Visual Feature Concatenation
Concatenation Layer

Fig. 3 Visual and textual feature extraction

right to left, depending on the information available. First, we tokenized the input question, padding it to the greatest length possible, and added the special token [CLS] to make it easier to understand. The [CLS] token is used in BERT for classification purposes, and it provides an aggregated representation of the entire sequence of characters it contains. Following that, we sent the tokenized sequence as an input to the BERT model (see Fig. 2). The [CLS] output of the BERT model was then fed into a neural network classifier, which was then used to classify the input question into one of the four categories that had been previously defined in the dataset.

3.2 Visual and textual feature extraction

The approach we utilized for feature extraction is described in detail in this section. In this study, a mixed deep learning model was used to recognize each picture feature. The model we propose is based on bidirectional long short-term memory, which may be used for text feature extraction (Bi-LSTM). A high-level overview of this technique is shown in Fig. 3, which emphasizes the following steps: (1) question

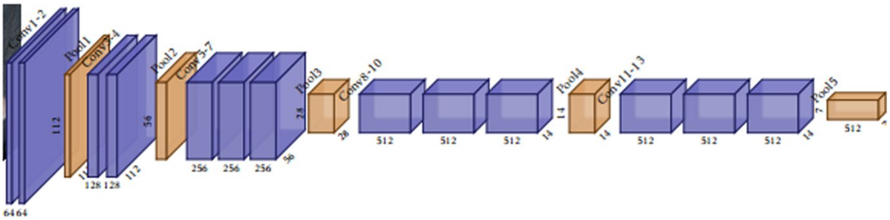

Fig. 4 VGG architecture

and picture pre-treatment; (2) image feature extraction; (3) text feature extraction; and (4) feature combination.

3.2.1 Visual feature extraction

In this section, we briefly describe the deep learning models used for feature extraction (VGG and ResNet).

A) VGG

The VGG model refines the architecture by proposing, among other things, to reduce the dimensions of the convolutions. Chatfield et al. [36] and Simonyan et al. [37] have suggested that it is simpler to optimize several successive convolutions of 3*3 kernels than a single convolution of 11*11 kernels. Furthermore, additional nonlinearities are likely to increase the expressiveness of the model.

The VGG model therefore replaces each classical wide convolution with a block of two or three successive 3*3 convolutions, as illustrated in Fig. 4.

B) ResNet

In 2015, He et al. [38] achieved an object recognition error rate of only 3.5% during the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Their approach consisted of a very deep network with over 100 convolutional layers. This optimization was possible partly because of batch normalization, but mostly because of residual learning.

The idea is to break the purely sequential structure of forward propagation networks by adding connections that short-circuit the next layer. These connections, called residuals, correspond to a simple identity operation and allow the activations and the gradient to traverse the whole network without suffering from evanescence or explosion due to the chain derivation rule.

The introduction of residual learning partially changed the paradigm previously used in the design of convolutional neural networks (CNN). The basic block constituting the network thus passes to the residual block. ResNet has many layers but comparatively few parameters, because only the last layer is fully connected (Fig. 5).

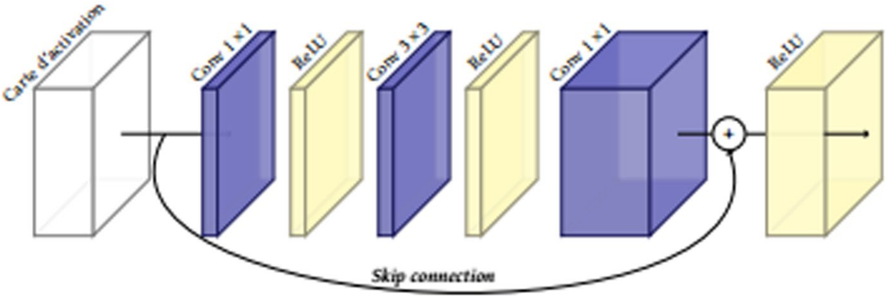

Fig. 5 Block convolution residual

3.2.2 Textual feature extraction

In this section, we describe the recurrent neural network (RNN) model used for textual feature extraction. Specifically, we used Bi-LSTM like an advanced RNN model to extract textual features due to the classic RNN model's inability to memorize.

RNNs are not effective in applications involving long time intervals, because their short-term memory is insufficient. Classic RNNs cannot memorize in the long term, and they forget after about 50 iterations. At the end of the nineties, researchers addressed these problems by developing more efficient methods, such as LSTM networks. These have notably revolutionized voice recognition by machines, as well as the understanding and generation of text messages.

LSTM networks are composed of a memory cell and, in most cases, a layer of neurons, in addition to three gates (input, output and forgetting). The sigmoid activation function of these gates allows them to influence the flow of information to and from the input, output and memory in an analog manner [39].

To improve the efficacy of LSTM, Song et al. [40] proposed the BLSTM model, which allows the use of information before and after the data is studied by the network at time t. The difference between the two architectures is shown in Fig. 6.

4 Pertinent answer retrieval

Multi-layer neural networks are arranged in three layers: a first layer connected to the outside of the network, or more outside the network; one or more hidden layers connected sequentially from the input layer; and an output layer.

In this step, we focused on the last layer to retrieve the optimal answer. For that, we combined the features extracted during the step described above and used them as input for our classifier based on a softmax layer and then determined the most accurate answer.

The activation function of the output layer is different from that of the hidden layers. The role of each layer different, as is its implementation. The last layer

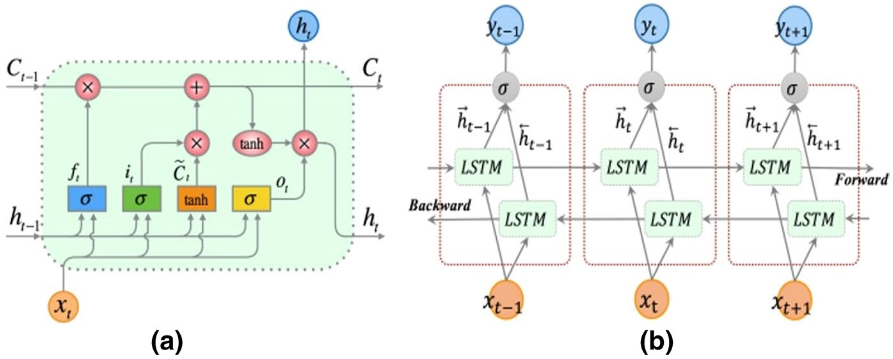

Fig. 6 **a** LSTM architecture. The pink circles are arithmetic operators and the colored rectangles are the gates in LSTM; **b** Unfolded architecture of Bidirectional LSTM [41]

for a classification task will enable the production of class probabilities for the input samples.

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{1}$$

which is a vector whose members are all x_i values, and which may take on any real value. All output values of the function must total to 1; hence, the bottom term is the normalization term, which assures that the probability distribution is legitimate.

5 Experiment results and discussion

Using the model described in this research, we conducted a series of experiments. Comparing the outcomes of various deep learning methods is also part of our research. An Rtx 2060 graphics card and 16 GB of RAM were used to create this model in Python.

Deep learning models were used in a series of trials before our approach could be used. There are four possible scenarios in our model:

1. *Scenario 1* Classification of medical questions using a deep learning model
2. *Scenario 2* VQA using one deep learning model for image feature extraction, and without the question classification step
3. *Scenario 3* VQA using a hybrid deep learning model based on an optimization algorithm, specifically the ADAM algorithm
4. *Scenario 4* investigates the effect of the Adam, AdaGrad, Stochastic gradient descent and RMS Prop optimization approaches on the network's performance

Table 1 ImageCLEF 2019 Dataset Description [42]

	ImageClef 2019 Dataset		
	Training	Validation	Test
Images	3200	500	500
Questions	12792	2000	500

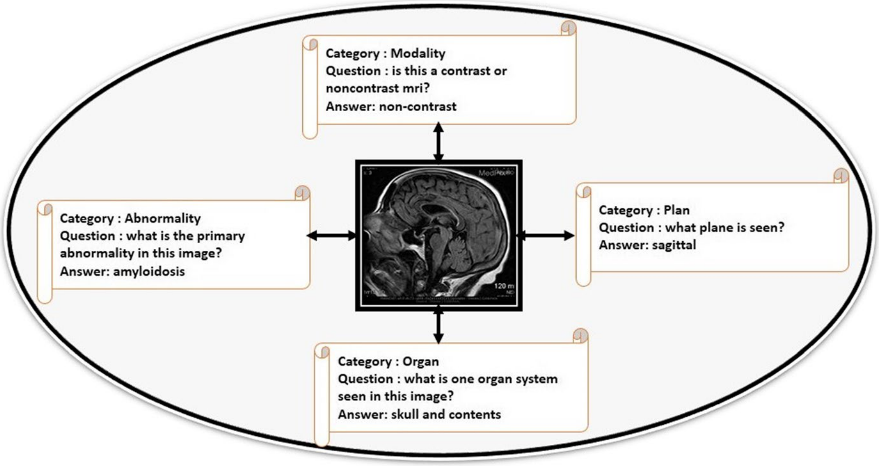
Category : Modality
Question : is this a contrast or noncontrast mri?
Answer: non-contrast
Category : Abnormality
Question : what is the primary abnormality in this image?
Answer: amyloidosis
Category : Plan
Question : what plane is seen?
Answer: sagittal
Category : Organ
Question : what is one organ system seen in this image?
Answer: skull and contents

Fig. 7 Sample from the ImageCLEF dataset comprising an image and a four-category question

5.1 Data description and evaluation metrics

The ImageClef 2019 [42] dataset was utilized to assess the performance of our improved deep learning model. Table 1 describes the dataset in greater detail.

There are four sorts of questions that may be asked about each dataset's picture: plan, modality, organ and anomaly. Images and four-category questions taken from the ImageCLEF data collection are shown in Fig. 7.

To evaluate the efficiency of the proposed model, we used the standard metrics employed in large-scale medical image sets for the visual QA tasks, such as BLEU, word-based semantic similarity (WBSS) and accuracy.

A system-generated answer and the real-world answer were compared using the BLEU [43] measure. To determine semantic similarity in the biomedical area, the second metric (WBSS) [44] was developed recently using Wu and Palmer Similarity (WUPS) [45] and WordNet ontology as a backend. A final metric from the general-domain VQA assignment is the accuracy metric, which measures the precise matches between a given response and the ground-truth answer.

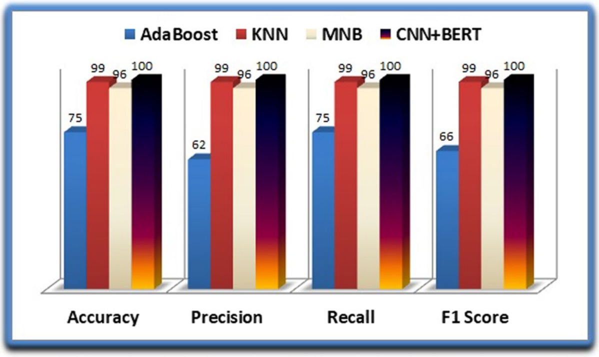

Fig. 8 Tests on alternative classifiers against the proposed CNN-BERT model

5.2 Evaluation of the classification method

Though the main idea of our approach was to implement a model based on a hybrid deep learning model, in this paper we chose to start by classifying medical questions into four categories. To that end, we compared several algorithms belonging to the traditional machine learning family. The algorithms tested in this category were K-Nearest Neighbors, Multinomial Naive Bayes and AdaBoost. To evaluate our classification method, we used standard metrics, such as accuracy, precision, recall and F-measure.

$$Precision = TP/(TP + FP) \tag{2}$$

$$Recall = TP/(TP + FN)(3) \tag{3}$$

$$F1 - score = 2/(1/P + 1/R)(4) \tag{4}$$

where

- *TP* true positive,
- *FP* false positive,
- *P* precision,
- *R* recall,

In Fig. 8, we show the results for all of the previously mentioned algorithms. Based on accuracy, the hybrid model CNN-BERT gave the best score, followed by the KNN algorithm.

We notice that the precision rate for the classification is very high. As a result, we investigated the database questions. In the end, we notice that questions from each category are very different when it comes to key words, especially after they've been pre-processed and stop words have been taken out. that the key words that remain are very significant for each category among the four. For example, we notice that for the modality category in most of the questions, we find a word among these three: ”modality, contrast, weighted”; on the other hand, at the level of questions in the plan category, we find the keyword ”plan” very frequently. In the category ”organ”, the most frequent keyword is ”organ”. This internal investigation and the use of a very relevant classifier for text like BERT, allowed us to conclude and validate these results, which are very high. Also, we may explain BERT’s excellent performance by pointing to the difficult contextual link between the keywords for the questions in each category, which is supported by the fact that BERT makes use of a transformer (an attention mechanism that learns the contextual relationships between words or subwords in a text).

5.3 Evaluation of the hybrid VQA-model

5.3.1 Effectiveness of the question classification method for VQA model

This section investigates the effectiveness of the classification step and the hybrid method in visual feature extraction. Our experiments consisted of the following steps:

- We used only one deep learning model for visual feature extraction and a classic textual feature such as TF/IDF, and without the question classification step (BaseLine)
- We used one deep learning model for visual feature extraction and a Bi-LSTM model for textual feature extraction, along with the question classification step (BaseBi-LSTM/BERT).
- We used our hybrid approach for feature extraction, along with the question classification step (hybrid approach/BERT).

Table 2 shows the comparison between the different runs, with the improvement rates of our runs compared to the runs that used BLEU, WBSS and accuracy. The experiments were performed using the ImageCLEF 2019 dataset. The best results are presented in bold. By comparing the four runs, we found that the best results were obtained when the VQA model included a classification step. Extracting and combining visual features using two deep learning models increased the accuracy of the answers.

In ImageClef2019 dataset, a question can be classified into one of four categories: abnormality; modality; organ system; and plan.

- *Abnormality* the questions are mostly divided into two categories: 1. an investigation into the presence of anomalies in the image and 2. an investigation into the nature of the abnormalities.
- *Modality* Inquiry into the sorts of medical pictures, such as MRIs and CT scans, that are available.
- *Organ* a question about which organ is seen in the photograph.

Table 2 Performance evaluation results of proposal models based on the ADAM optimization algorithms

	ImageClef 2019 Dataset											
	Modality			Plan			Organ			Abnormality		
	BLEU	WBSS	Acc	BLEU	WBSS	Acc	BLEU	WBSS	Acc	BLEU	WBSS	Acc
BaseLine	0.18	0.2	0.16	0.2	0.22	0.2	0.19	0.21	0.18	0.02	0.024	0.019
BaseBi-LSTM	0.2	0.22	0.19	0.21	0.21	0.21	0.2	0.22	0.2	0.03	0.032	0.03
BERT/BaseBi-LSTM	0.4	0.43	0.39	0.46	0.48	0.42	0.39	0.4	0.39	0.04	0.042	0.039
BERT/Hybrid approach	**0.48**	**0.45**	**0.43**	**0.52**	**0.56**	**0.52**	**0.45**	**0.41**	**0.41**	**0.056**	**0.096**	**0.056**

Table 3 BLEU rate of hybrid proposed models based on the applied optimization algorithms for ImageCLEF 2019 dataset

	Modality	Plan	Organ	Abnormality
ADAM	0.48	0.52	0.45	0.051
AdaGrad	0.37	0.42	0.38	0.046
SGD	0.45	0.52	0.41	0.056
RMSProp	0.35	0.38	0.32	0.02
Mean	0.41	0.46	0.39	0.043

- *Plan* an investigation into the captured plan, such as whether it is vertical or horizontal.

There are 1,461 distinct abnormalities in the 3,082 training photos for Type 1 questions, and 407 different abnormalities in the 477 validation images. When it comes to the category of abnormality, the results we obtained are considerably different from those obtained for the other category. This is due to the fact that the training set contains a small number of answers for a large number of different abnormalities.

5.3.2 Impact of the optimization algorithm in the deep learning model

One of the ultimate goals of this study was to get the best prediction scores from the models that was created. When the models are built in line with their architectural frameworks, certain parameter values are critical to boosting prediction success. In deep learning and machine learning applications, loss is defined as the difference between predicted and actual values, and minimizing loss implies strong model performance. The model's loss function must be minimized in order to keep losses to a minimum. Optimization methods such as Adam [46], AdaGrad [47], SGD and RMSProp [48] may be used to tackle this issue. The performance of each method in obtaining the global minimum, which is the lowest loss value, may differ. It has therefore been a primary research issue to investigate the impact of these algorithms on the three models generated. As a result, these four solutions have been tested as network-optimizers.

To train deep neural networks, the learning rate and batch size parameters were adjusted to 0.0001 and 32, respectively, in experiments using both the proposed CNN architecture and transfer learning approaches. A total of 100 epochs were used in the trials. Experiments take longer to complete when the epoch value grows.

In terms of the suggested model's mean accuracy, Adam achieved an accuracy of 0.52 for plan category and 0.48 for modality category, whereas SGD approach achieved an accuracy of 0.45 and 0.52, respectively, in modality and plan category, as shown in Table 3. Adam and SGD optimization techniques were shown to perform better in this study's tests, than the other two models.

While Adam was used to get the best accuracy in the modality category, SGD was used to produce the best results in abnormality category that used the proposed

Table 4 A comparison of the suggested model run's performance to that of the official ImageClef 2019 runs [42]

	VQA-Med 2019: Accuracy scores			
	Modality	Plan	Organ	Abnormality
Proposal Model	**0.48**	**0.52**	**0.45**	**0.056**
Hanlin	0.202	0.192	0.184	**0.046**
yan	0.202	0.192	0.184	0.042
minhvu	0.210	0.194	0.190	0.022
TUA1	0.186	0.204	0.198	0.018
UMMS	0.168	0.190	0.184	0.02
AIOZ	0.182	0.180	0.182	0.020

Bold values indicate the best results obtained by the different techniques presented in this table for each category

approach. Adam and SGD were shown to be more effective in the experiments carried out.

5.4 Proposed model compared to ImageClef 2019 official submissions without question classification step

For the purpose of demonstrating the efficacy of our proposed approach, we compared our findings to those acquired by official contributors to the ImageClef 2019 [42]. Knowing that the findings provided in the Table 4 were obtained prior to the categorizing of the questions into categories is comforting. Following the official relevance evaluations conducted by ImageClef 2019 in three areas (plan, organ and modality), our run would be rated first in all three categories, as indicated in Table 4. Based on the findings of this comparison, we can conclude that the classification of questions has a direct impact on the ability of our model to predict the correct answer. To be more explicit, by dividing inquiries into categories, our model is able to determine the purpose that lies behind users' requests. Also, the results generated by our hybrid model without the classification phase demonstrate the usefulness of increasing the amount of characteristics that are represented by each individual picture.

The most accurate performance for the plan, organ and modality categories was obtained by our model, thanks to the hybrid model's increased number of features.

6 Conclusion

This paper presented a new hybrid deep learning model for a visual question/answer system designed for the medical field. Our technique for visual feature extraction is based on a hybridization of two deep learning models for feature extraction. First, medical questions are categorized using a new technique to text categorization that is applied to the questions themselves. Based on the combination of a BERT model with a CNN classifier, this model has been developed. In order to extract textual features from the queries, a bidirectional long short-term memory (Bi-LSTM) model for textual feature extraction was used. The second input was presented as an image and was mapped in two different deep learning models for visual feature extraction. Before any answers were generated, we combined all of the visual and textual features and put them into a full layer based on the softmax layer. We evaluated our hybrid method using the ImageClef 2019 dataset and demonstrated the superiority of our method as compared to previous works.

This research's end goal was to obtain the best possible prediction scores from the models that were constructed. Optimization methods such as Adam, AdaGrad, SGD and RMSProp may be used to achieve this aim. While Adam was used to get the highest level of accuracy in the modality category, SGD was used to achieve the highest level of accuracy in the abnormality category when the proposed model was utilized. In the studies that were conducted, it was discovered that Adam and SGD were more effective.

Future work will focus on refining the question classification system. There is a semantic gap between the regional visual features and the text of the question. For that reason, we intend to develop the classification system by using an expansion method based on medical ontology.

References

1. He X, Cai Z, Wei W, Zhang Y, Mou L, Xing E, Xie P (2021) Towards visual question answering on pathology images. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 708–718
2. Demner-Fushman D, Lin JJ (2006) Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In: ACL
3. Lin JJ, Katz B (2003) Question answering from the web using knowledge annotation and knowledge mining techniques. In: CIKM '03
4. Popescu A-M, Etzioni O, Kautz HA (2003) Towards a theory of natural language interfaces to databases. In: IUI '03
5. Rinaldi F, Dowdall J, Schneider G, Persidis A (2004) Answering questions in the genomics domain. In: ACL 2004
6. Katz B (1997) From sentence processing to information access on the world wide web. In: AAAI Spring Symposium on Natural Language Processing for the World Wide Web, vol. 1, p. 997
7. Lin Z, Zhang D, Tac Q, Shi D, Haffari G, Wu Q, He M, Ge Z (2021) Medical visual question answering: a survey. arXiv preprint arXiv:2111.10056
8. Devlin J, Chang M-W, Lee K, Toutanova K (2019) Bert: pre-training of deep bidirectional transformers for language understanding. In: NAACL
9. Lehnert WG (1977) Human and computational question answering. Cogn Sci 1:47–73
10. Do T, Nguyen BX, Tjiputra E, Tran M, Tran QD, Nguyen A (2021) Multiple meta-model quantifying for medical visual question answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 64–74. Springer
11. Liu B, Zhan L-M, Xu L, Ma L, Yang Y, Wu X-M (2021) Slake: a semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1650–1654. IEEE
12. Gassara A, Rodriguez IB, Jmaiel M, Drira K (2017) A bigraphical multi-scale modeling methodology for system of systems. Comput Electr Eng 58:113–125

13. Monceaux L, Robba I (2002) Les analyseurs syntaxiques : atouts pour une analyse des questions dans un système de question-réponse ? In: JEPTALNRECITAL
14. Mendes S, Moriceau V (2004) L'analyse des questions: intérêts pour la génération des réponses. In: Workshop Question-Response
15. Ferret O, Grau B, Hurault-Plantet M, Illouz G, Jacquemin C, Masson N, Lecuyer P (2000) Qalc–the question-answering system of limsi-cnrs. In: TREC
16. Grau B, Ligozat A-L, Robba I, Vilnat A, Monceaux L (2006) Frasques: a question-answering system in the equer evaluation campaign. In: LREC 2006, p. 2006
17. Laurent D, Séguéla P (2005) Qristal, système de questions-réponses. In: Actes de la 12ème Conférence sur Le Traitement Automatique des Langues Naturelles. Articles longs, pp. 51–60
18. Benamara F (2004) Cooperative question answering in restricted domains: the webcoop experiment. In: Proceedings of the Conference on Question Answering in Restricted Domains, pp. 31–38
19. Teillaud JS (2017) medecine/sciences 2017: the french touch des avancées des connaissances biomédicales en... langue française. M S-Med Sci 33:7–8
20. Zweigenbaum P (2001) Traitements automatiques de la terminologie médicale. Revue française de linguistique appliquée 6(2):47–62
21. Khabou N, Rodriguez IB (2015) Threshold-based context analysis approach for ubiquitous systems. Concurr Comput Pract Exp 27(6):1378–1390
22. Malinowski M, Fritz M (2014) A multi-world approach to question answering about real-world scenes based on uncertain input. In: NIPS
23. Agrawal A, Lu J, Antol S, Mitchell M, Zitnick CL, Parikh D, Batra D (2015) Vqa: visual question answering. Int J Comput Vision 123:4–31
24. Goyal, Y, Khot, T, Summers-Stay, D, Batra, D, Parikh, D. (2017) Making the v in vqa matter: Elevating the role of image understanding in visual question answering. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6325–6334
25. Frome, A, Corrado, G.S, Shlens, J, Bengio, S, Dean, J, Ranzato, M, Mikolov, T.: Devise: a deep visual-semantic embedding model. In: NIPS (2013)
26. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Commun ACM 60:84–90
27. Mikolov, T, Chen, K, Corrado, G.S, Dean, J.: Efficient estimation of word representations in vector space. In: ICLR (2013)
28. Wang L, Li Y, Huang J, Lazebnik S (2019) Learning two-branch neural networks for image-text matching tasks. IEEE Trans Pattern Anal Mach Intell 41:394–407
29. Huang, Y, Wang, W, Wang, L.: Instance-aware image and sentence matching with selective multimodal lstm. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7254–7262 (2017)
30. Dong, X, Zhu, L, Zhang, D, Yang, Y, Wu, F (2018): Fast parameter adaptation for few-shot image captioning and visual question answering. Proceedings of the 26th ACM international conference on Multimedia
31. Mao J, Gan C, Kohli P, Tenenbaum JB, Wu J (2019) The neuro-symbolic concept learner: interpreting scenes, words, and sentences from natural supervision. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA
32. Yi K, Wu J, Gan C, Torralba A, Kohli P, Tenenbaum JB (2018) Neural-symbolic vqa: disentangling reasoning from vision and language understanding. In: NeurIPS
33. Nguyen BD, Do T-T, Nguyen BX, Do TK, Tjiputra E, Tran QD (2019) Overcoming data limitation in medical visual question answering. In: MICCAI
34. Masci J, Meier U, Ciresan DC, Schmidhuber J (2011) Stacked convolutional auto-encoders for hierarchical feature extraction. In: ICANN
35. Finn C, Abbeel P, Levine S (2017) Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML
36. Chatfield K, Simonyan K, Vedaldi A, Zisserman A (2014) Return of the devil in the details: delving deep into convolutional nets. In: Valstar MF, French AP, Pridmore TP (eds) British Machine Vision Conference, BMVC 2014. Nottingham, UK
37. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Bengio, Y, LeCun, Y (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, Conference Track Proceedings
38. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 16 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778

39. Ghourabi A, Mahmood MA, Alzubi QM (2020) A hybrid cnn-lstm model for sms spam detection in Arabic and English messages. Future Internet 12:156
40. Song M, Zhao X, Liu Y, Zhao Z (2018) Text sentiment analysis based on convolutional neural network and bidirectional lstm model. In: ICPCSEE
41. Cui Z, Ke R, Pu Z, Wang Y (2020) Stacked bidirectional and unidirectional lstm recurrent neural network for forecasting network-wide traffic state with missing values. ArXiv
42. Abacha AB, Hasan SA, Datla V, Liu J, Demner-Fushman D, Müller H (2019) Vqa-med: overview of the medical visual question answering task at imageclef 2019. In: CLEF
43. Papineni K, Roukos S, Ward T, Zhu W-J (2002) Bleu: a method for automatic evaluation of machine translation. In: ACL
44. Sogancioglu G, Öztürk H, Özgür A (2017) Biosses: a semantic sentence similarity estimation system for the biomedical domain. Bioinformatics 33:49–58
45. Wu Z, Palmer MS (1994) Verb semantics and lexical selection. In: 32nd Annual Meeting of the Association for Computational Linguistics, New Mexico State University, pp. 133–138
46. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: Bengio Y, LeCun Y. (Eds) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings
47. Duchi JC, Hazan E, Singer Y (2011) Adaptive subgradient methods for online learning and stochastic optimization. J Mach Learn Res
48. Tieleman T, Hinton G (2012) Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. University of Toronto, Technical Report

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.