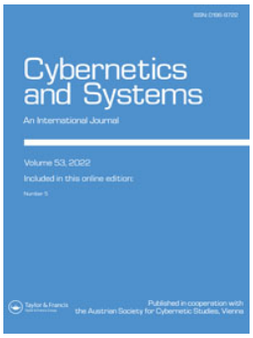

Taylor & Francis
Taylor & Francis Group

Cybernetics and Systems
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucbs20

Optimal Deep Neural Network-Based Model for Answering Visual Medical Question

Karim Gasmi, Ibtihel Ben Ltaifa, Gaël Lejeune, Hamoud Alshammari, Lassaad Ben Ammar & Mahmood A. Mahmood

To cite this article: Karim Gasmi, Ibtihel Ben Ltaifa, Gaël Lejeune, Hamoud Alshammari, Lassaad Ben Ammar & Mahmood A. Mahmood (2022) Optimal Deep Neural Network-Based Model for Answering Visual Medical Question, Cybernetics and Systems, 53:5, 403-424, DOI: 10.1080/01969722.2021.2018543

To link to this article: https://doi.org/10.1080/01969722.2021.2018543

CYBERNETICS AND SYSTEMS: AN INTERNATIONAL JOURNAL
2022, VOL. 53, NO. 5, 403–424
https://doi.org/10.1080/01969722.2021.2018543

Optimal Deep Neural Network-Based Model for Answering Visual Medical Question

Karim Gasmi[a], Ibtihel Ben Ltaifa[b], Gaël Lejeune[b], Hamoud Alshammari[c], Lassaad Ben Ammar[d], and Mahmood A. Mahmood[c]

[a]Department of Computer Science, College of Arts and Sciences at Tabarjal, Jouf University, Jouf, Saudi Arabia; [b]STIH, Sorbonne Université, Paris, France; [c]Department of Information Systems, College of Computer and Information Sciences, Jouf University, Jouf, Saudi Arabia; [d]College of Sciences and Humanities, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia

ABSTRACT

Over the last few years, the amount of available information has increased exponentially in all professional fields, including the medical field. Modern-day patients have access to a wealth of medical information, whether it be from brochures, newspapers, television campaigns, or internet documents. To facilitate and accelerate the search for medical information, more precise systems have been implemented, such as visual question-and-answer systems. A visual question-and-answer system is designed to provide direct and precise answers to questions asked in natural language. In this context, we propose an optimal deep neural network model based on an adaptive optimization algorithm, which takes medical images and natural language questions as input, then provides precise answers as output. Our model starts by classifying medical questions following an embedding phase. We then use a deep learning model for visual and textual feature extraction and emergence. In this paper, we aim to maximize the accuracy rate and minimize the number of epochs in order to accelerate the process. This is a multi-objective optimization problem. The selection of deep learning model parameters, such as epoch number and batch size, is an essential step in improving the model, thus, we use an adaptive genetic algorithm to determine the optimal deep learning parameters. Finally, we propose a dense layer for answer retrieval. To evaluate our model, we used the ImageCLEF 2019 VQA data set. Our model outperforms existing visual question-and-answer systems and offers a significantly higher retrieval accuracy rate.

KEYWORDS

Bi-LSTM; deep learning; EfficientNet; genetic algorithm; medical visual question answering; optimization

Introduction

Recent research has focused on merging natural language processing and computer vision tasks with the goal of improving the interactions between humans and intelligent systems, which in turn has given rise to such tasks

CONTACT Karim Gasmi kgasmi@ju.edu.sa Department of Computer Science, College of Arts and Sciences at Tabarjal 74769, Jouf University, Jouf, Saudi Arabia; Ibtihel Ben Ltaifa ibtihel.benltaifa@sorbonne STIH, Sorbonne Université, Paris, France.

as visual question answering (VQA), visual question generation (VQG), and multimodal machine translation, among others (Al-Sadi, Al-Theiabat, and Al-Ayyoub 2020).

VQA has gained a lot of attention from academics and researchers due to its combination of computer vision, natural language processing (NLP), and artificial intelligence (AI). Given an image and a natural language question relating to the image, a VQA system will try to produce a correct natural language answer (Agrawal et al. 2017). The broader concept is to design systems that can understand the contents of a given image similar to the way humans do, and that can correctly answer questions about that image in natural language.

This is a challenging task, because it requires a broad knowledge of image processing, NLP, and multimodal learning from both visual and linguistic data. VQA systems combine NLP, which provides an understanding of the question and the ability to produce an answer, with computer vision techniques, which provide an understanding of the content of the image.

In recent years, AI technology has come to dominate the healthcare service system and is still growing, with early successes involving ChatBots, diagnostic tools, and radiological image analysis (Emilio 2020). These emerging technologies have replaced the traditional system with AI systems that automatically carry out human medical processes.

VQA shows great potential in interpreting automated medical imagery (e.g., radiology images) and machine-supported diagnoses, and can therefore assist with clinical decisions (Liao et al. 2020). Since 2016, a general VQA challenge has been issued every year to answer questions of various types (i.e., multiple choice questions, yes/no questions, and open-ended questions). In 2018, ImageCLEF launched the VQA-Med challenge, which is specific to the medical field and is held annually (Ionescu et al. 2020).

For VQA-Med, it is necessary to combine visual features from images with textual features from questions. A network cannot produce a correct answer unless it understands what the questions are asking and which image features will provide the answer. Thus, it is crucial to build an efficient system to embed images and text. A key solution to VQA is the fusion of visual and textual features extracted from input images and questions.

The selection of the most efficient neural network model is delicate because it depends on many factors, especially the input training data, the feature extraction or selection method, and the chosen hyperparameters. Given the limited understanding of neural networks trained for tasks as complex as VQA, even slight variations in hyperparameters and network architectures may have significant and sometimes unpredictable effects on final performance. Hyperparameter optimization is a critical part of any

deep learning process. A good deep learning model is not enough to achieve an exceptional performance; it needs to be fine-tuned through neural network training.

For these purposes, we propose an optimal deep neural network-based model for medical VQA. The proposed model uses an adaptive optimization algorithm that takes medical images and natural language questions as input, and provides precise answers as output. The goal of this work is to identify the best hyperparameters for deep learning architectures to achieve exceptional performances. Specific contents within a given image must be interpreted as indicated by the linguistic context of the question in order to generate accurate answers. To evaluate our model, we used the ImageCLEF 2019 VQA data set. The main contributions of this paper are as follows:

- To organize the questions from the data set, we propose a classification process to assign questions to relevant categories using the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model.
- We used different deep learning models for visual and textual feature extraction and emergence. We first used the EfficientNet model to extract visual features. To better capture the relationships between words that are relatively far apart and improve the extracted semantics, we used a bidirectional long short-term memory (Bi-LSTM) model to encode the questions. We then combined the encoded questions with the image features using an attention model.
- For optimization purposes, our goal was to maximize the accuracy rate and minimize the number of epochs in order to accelerate the process. This is a multi-objective optimization problem. The selection of deep learning model hyperparameters, such as epoch number and batch size, is an essential step in improving the model; thus, we used an adaptive genetic algorithm to determine the optimal deep learning parameters.

The rest of this paper is organized as follows: in Section 2, we review recent relevant innovations and trends in medical VQA. Section 3 presents the architecture of our deep learning model. Section 4 details the experimental evaluation of our model and discusses the results in comparison to other works, with corresponding analyses. Conclusions and future works are discussed in Section 5.

Related Work on Visual Question Answering

Several works have studied visual question answering (VQA) tasks. These works can be divided into two categories: (1) general VQA and (2) medical VQA. The main difference between medical and general VQA is the size of

the data set. General VQA can create a massive data set because external knowledge or common-sense knowledge is sufficient to generate questions and answers. Medical VQA, on the other hand, requires medical experts, which poses great difficulties in collecting enough data to generate accurate responses. Medical imagery, such as radiology images, are field-specific, and can only be interpreted by well-trained medical experts.

For general VQA, several researchers have merged visual and textual information using a convolutional neural network (CNN) and recurrent neural network (RNN) encoder-decoder formulation (Malinowski, Rohrbach, and Fritz 2015), while others placed attention mechanisms over images to highlight the most important regions in the images, making it easier to predict the right answers (Shih, Singh, and Hoiem 2016). In Lu et al. (2016), the authors proposed a co-attention mechanism that simultaneously focuses on images and text. During the co-attention process, the image representation is used to guide the question attention mechanism, and the text representation is used to guide image attention mechanism.

Medical VQA Literature Review

In this work, we studied VQA in the medical field. Medical VQA is challenging, as it requires a dedicated medical data set as well as medical experts who can understand and interpret the data. Several methods have been proposed in medical VQA literature, and can be divided into three categories based on the nature of their main contribution: (1) embedding approaches based on deep neural networks; (2) attention mechanisms; and (3) models based on external knowledge bases.

Embedding Approaches

VQA deep learning methods receive widespread attention due to their performance on various computer vision and natural language processing (NLP) tasks (Ioannidou et al. 2017; Torfi et al. 2020). Several works proposed deep learning approaches, also known as embedding approaches. Embedding approaches use convolutional and recurrent neural networks (CNNs and RNNs) to learn representations of images and sentences in a common feature space. A representation in a common space enables learning interactions and can help researchers make inferences between the questions and the images (Wu et al. 2017). Image features are produced by a CNN pre-trained for object recognition, and text features are obtained with word embeddings pre-trained on long texts using RNNs with long short-term memory (LSTM) (Malinowski, Rohrbach, and Fritz 2015; Ren, Kiros, and Zemel 2015).

The authors in Malinowski, Rohrbach, and Fritz (2015) proposed an approach that combines a CNN with an LSTM, thereby creating an end-to-end architecture that predicts answers. An RNN is designed to handle questions and answers of different sizes. For that purpose, the authors used common shared weights between the encoder and decoder LSTMs, whereas the authors in Gao et al. (2015) proposed a multimodal QA method, which uses LSTMs to encode questions and produce answers. The multimodal method learns distinct parameters and only shares the word embedding.

Additionally, while the authors in Malinowski, Rohrbach, and Fritz (2015) fed CNN pre-trained image features and questions to the encoder LSTM together, CNN features in Gao et al. (2015) were not fed into the encoder prior to the question, but at every time step. The authors in Ren, Kiros, and Zemel (2015) proposed other technical improvements with a Bi-LSTM model to encode and decode sentences. This model treats each given image as one word of the question, and only produces the answer during the last time step. The authors demonstrated that Bi-LSTMs are more effective in capturing relationships between distantly spaced words within the questions.

The authors in Noh, Seo, and Han (2016) proposed a novel architecture for VQA based on two subnetworks: (1) the classification network and (2) the parameter prediction network. The classification network used CNN with a dynamic parameter layer, where its weights are determined based on the question. For adaptive parameter prediction, the authors employed gated recurrent units (GRUs), a variant of LSTMs. This arrangement significantly improved answer accuracy, as compared to Malinowski, Rohrbach, and Fritz (2015) and Ren, Kiros, and Zemel (2015). The authors in Saito et al. (2017) took advantage of the discriminative information provided by the image and text features by performing separate operations (addition and multiplication) for input features to form a common embedding space. The proposed method performed well even without an attention mechanism. In Kafle and Kanan (2016), the authors proposed an approach that can predict the form of the answer based on the question and then formulate the answer using a Bayesian framework.

Other VQA approaches did not use RNNs to encode questions. The authors in Ma, Lu, and Li (2016) used a CNN model to learn not only the image and question representations, but also their interactions and the relationships between the images and questions, to predict the answers. The authors in Zhou et al. (2015) and Agrawal et al. (2017) both used a traditional bag-of-words model for the questions.

Attention-Based VQA Model

When looking at an image, the viewer's focus is necessarily on a specific part of the image, rather than on the whole image (He and Han 2020).

Most of the proposed VQA approaches described above use global image features to represent the visual input, which can result in irrelevant or noisy information during the prediction stage. To address this issue, some researchers developed a series of methods based on attention mechanisms. One method, based on image-guided attention mechanisms, uses local image features and allows the model to assign different weights to features from different regions.

The authors in Lu et al. (2016) proposed a hierarchical co-attention mechanism for VQA. The co-attention architecture represented different regions of the images and different fragments of the questions. It used combination and average pooling to merge all components. The authors in Shih, Singh, and Hoiem (2016) proposed an image-region selection-based model that learned to identify which image regions were relevant to which questions. In Fukui et al. (2016), the authors proposed a multimodal pooling-based method to combine visual and textual features. They integrated the soft attention mechanism into the multimodal pooling method. More specifically, they used two convolutional layers to predict the weight of the attention mechanism for each location, based on both the visual and textual representations. They then took a weighted sum of the spatial vectors, using the attention map to create the visual representation with its attention mechanism. By allowing the model to learn to focus on salient locations based on both features, they consistently showed the benefits of combining multimodal features and of predicting attention maps with the multimodal pooling method.

The authors in Li and Jia (2016) proposed an end-to-end trainable neural network-based model that updated the question representation iteratively by selecting image region information through multiple reasoning layers. They employed an attention mechanism to generate the attention distribution over those image regions that closely related to the questions. The authors in Teney et al. (2018) proposed a method that detected image features using bottom-up and top-down attention. The method was trained to focus on specific elements and regions in a given image using Visual Genome annotations (Krishna et al. 2017).

In Kim, Jun, and Zhang (2018), the authors proposed a bilinear attention network that extended unitary attention networks. It used bilinear attention maps, where the joint representations of multimodal inputs were extracted using the low-rank bilinear pooling technique. The authors in Nguyen and Okatani (2018) proposed a dense co-attention network architecture to improve the fusion of visual and textual representations. The core of the proposed network was the dense co-attention layer, which was designed to form a hierarchy for multi-step interactions between an image-question pair. The authors in Yu et al. (2019) proposed a deep modular co-attention

network that consisted of modular co-attention layers cascading in depth. Each layer modeled the self-weighting of questions and images by associating the keywords in the questions with the critical regions in the images.

Models using External Knowledge Bases

Most existing VQA data sets largely comprise questions that require common sense and little prior or basic factual knowledge. Thus, performances using these data sets poorly reflect the particular capabilities of the models built upon them. Knowledge-based VQA does not have access to enough image content to answer questions, encouraging methods that rely on external knowledge resources.

Several VQA studies proposed to take advantage of the benefits of external knowledge bases. A knowledge base can be viewed as a large-scale graph structure that connects different entities with their relations (Richardson and Domingos 2006). Large-scale knowledge bases are constructed either by manual annotation or crowd-sourcing, as with DBpedia (Lehmann et al. 2015), Freebase (Dalton, Dietz, and Allan 2014), YAGO (Rebele et al. 2016), ConceptNet (Shen and Kejriwal 2020), and more. These databases store common-sense and factual knowledge provided by multiple sources. Linking such knowledge bases to VQA methods is an important factor in automated reasoning systems.

The authors in Marino et al. (2019) proposed a data set for questions that require external knowledge resources in order to be answered. The authors in Wang et al. (2017) proposed a VQA method that uses DBpedia as its source of external information. The proposed method is capable of forming conclusions on the contents of images and questions relative to the data in the knowledge base and the connections between them, as shown in the constructed knowledge graph. The main limitation of this method is that it can only handle certain types of questions.

The authors in Wang et al. (2018) proposed an approach that learned to map images and questions to queries using two additional knowledge bases, ConceptNet and WebChild. In Wu et al. (2016), the authors proposed an LSTM-based approach capable of deeper image analysis and even common sense, which they achieved using DBpedia as their knowledge base. The proposed approach allowed questions to be asked on the contents of an image, even when a given image did not contain the whole answer.

The quality of the information in the knowledge base is one of the primary issues in these approaches. Existing knowledge bases are constructed by analyzing Wikipedia, and, as with their source material, are inconsistent at best. Manually curated knowledge bases are of necessity very topic-specific. To address this issue, the authors in Zhu et al. (2015) proposed a scalable knowledge base construction system that performed learning and

inference on a large-scale multimodal knowledge base. In particular, they built a scalable knowledge base to answer a variety of visual queries without re-training.

Medical VQA Image CLEF Challenges

Several methods have been applied specifically in the medical field through the VQA challenges organized by ImageCLEF[1]. During ImageCLEF 2020, some works dealt with medical VQA and visual question generation (VQG) tasks within the same challenge (Al-Sadi, Al-Theiabat, and Al-Ayyoub 2020; Liao et al. 2020).

The authors in Al-Sadi, Al-Theiabat, and Al-Ayyoub (2020) proposed a set of models to perform medical VQA and VQG. All proposed models were based on pre-trained CNNs, a pre-trained VGG, and data augmentation. They showed that treating the tasks as image classification tasks is more useful than including an NLP component, as is currently done in VQA/VQG tasks. With the VQA-Med task, they demonstrated that applying data augmentation to the images improved the performance of the models, as compared to those without augmentation. They showed the positive effects of using various augmentation methods, such as rotation change, width/height shift, rescaling, zooming, and ZCA whitening. Following a series of experiments, results showed that the best proposed model is based on VGG16, and that it augments the data using ZCA whitening.

The authors in Liao et al. (2020) proposed a set of VQA and VQG approaches. The center of the proposed approach is a knowledge inference method we named Skeleton- based Sentence Mapping (SSM), which helped analyze the predicted answers. The SSM method was designed to determine question categories and assign corresponding labels. As a result, the authors turned the VQA problem into a multi-task image classification problem, which allowed us to focus on the imaging modality. They adopted a class-wise and task-wise normalization method to facilitate the optimization of multiple tasks with incomplete labels in a single network for robust prediction. Additionally, the authors in Sarrouti (2020) proposed a variational autoencoder-based method and a multi-class image classification-based method to address the VQA-Med problem. The variational autoencoders-based method takes a medical question-image pair as input and generates a natural language answer as output. The encoder consists of a pre-trained CNN model and LSTM to encode the dense vectors of the images and the questions, respectively, into a latent space. The decoder network also uses LSTM to decode questions from the latent space.

[1] https://www.imageclef.org/

Other works have focused only on medical VQA. The authors in Verma and Ramachandran (2020) proposed an encoder-decoder-based model where the encoder takes two inputs—a question and a radiology image—and the decoder generates the answer. The encoder first takes a set of feature vectors from images obtained using VGG16, and a set of attention features for each question using Bidirectional Encoder Representations from Transformers (BERT). These two extracted features are fused using an advanced feature fusion technique called multimodal factorized bilinear pooling (MFB) with question image fuze attention. The outputs from the encoder are thought vectors and encoder LSTM sequences. The decoder takes these outputs and tries to predict the answer word for word. The answers are encoded using Global Vectors (GloVe) word embeddings. This model achieved a high accuracy rate.

Due to the unequal amount of information in images and questions in medical VQA, the authors in Chen, Gong, and Li (2020) tried to implement effective visual representation. After deep analysis of the distribution of information in the training set, they concluded that the problem was the limited amount of training data. To resolve this issue, they designed a method of extending the training set using the Kullback-Leibler divergence with a cumulative learning strategy.

The authors in Mohamed and Srinivasan (2020) proposed an encoder-decoder model for medical VQA. Under the proposed model, the VGGNet and LSTM techniques are used to extract the image and text feature vectors during the encoder stage. The generated feature vectors are then combined and given to the decoder as inputs to predict the answer. For evaluation, the authors used the ImageCLEF VQA-Med 2020 data set (original data set) and two other data sets, which they modified (reduced data set and augmented reduced data set). Experimental results showed that VGGNet can effectively extract medical image features from a small data set.

Deep learning has had a transformative impact on VQA. Existing models and procedures used in VQA studies are so complex that it is often impossible to determine the impact of the individual design and engineering choices for each model. This ambiguity delays progress in the field of VQA. To address this problem, future studies should focus on the effects of neural network architecture and hyperparameters on model performance. The details of each architectural choice and hyperparameter should be carefully selected for optimal performance in medical VQA.

Optimal Deep Neural Network-Based Model

In this section, we describe our proposed deep neural network architecture-based model. For the sake of transparency, we also discuss the specific

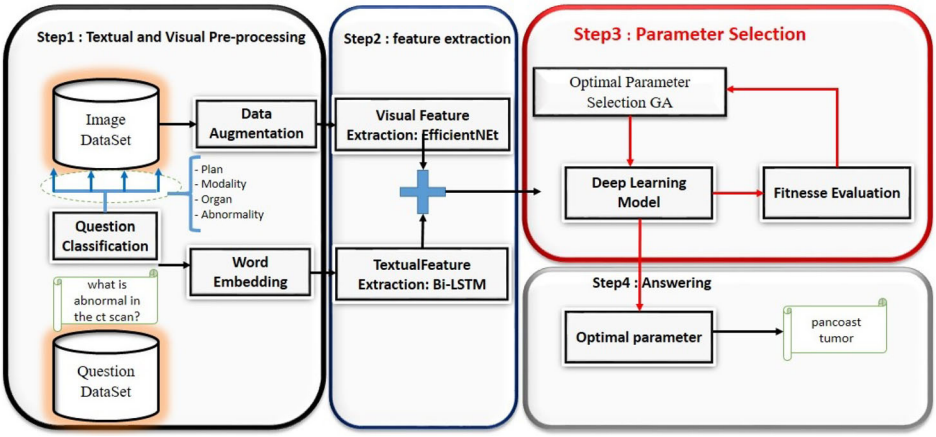

Figure 1. Proposal model based in hyperparameter deep learning selection.

choices and hyperparameter values that lead to its best performance. Section 4 will examine variations in the architecture and hyperparameters, and their influence on performance. The complete model is shown in Figure 1. It implements the well-known joint recurrent and convolutional neural network (RNN and CNN) embedding of questions and images, with the questions guiding the attention mechanisms over the images.

Pre-Processing

It is commonly known that too little training data results in a poorly trained neural network, and that any model using so little data will correspondingly be a poor approximation of reality. An overly constrained model will underfit the small training data set while an under-constrained model will likely overfit, both of which will result in a poor performance.

A neural network model is only as good as the data fed to it. Augmentation techniques make it possible to prevent the network from learning irrelevant patterns, boosting overall performance. Data augmentation encompasses a suite of techniques that enhance the size and quality of training data sets to build effective deep learning models. It is a very powerful method that has been developed to reduce overfitting, an issue that occurs in machine learning when the model does not extrapolate well from training data to invisible data.

In our work, we tried to reduce overfitting by using various image enhancement techniques. A total of 10 photos were taken for each image, which was then reconstructed using very small random rotations, offsets, scaling, and clipping. We used stemming and lemmatization to transform verbs, nouns, and other words to their original forms in order to prevent overfitting. We also eliminated all stop words and special characters. If

medical abbreviations coexisted with the original terminology, we attempted to convert them. Finally, to maximize training efficiency, we excluded low-frequency words from the data analysis. To improve results, each question can be assigned to one of four categories—plan, modality, organ, and abnormality—in the ImageCLEF 2019 data set. We suggest an approach based on the BERT model as a solution to this problem (Devlin et al. 2019).

Question Classification

The goal of this step is to divide the data set questions into four distinct categories: modality, plan, organ, and abnormality. The Bidirectional Encoder Representations from Transformers (BERT) pre-trained model was employed for this purpose (Devlin et al. 2019).

Since it was debuted by Google in 2018, BERT has been a huge hit. BERT, along with other transformer-based models, has had a significant impact on natural language processing (NLP). NLP tasks, such as question answering, sentiment analysis, text classification, and machine translation (Ghourabi 2021), were able to provide cutting-edge findings and high-performance models because of this technology. One of its advantages is its ability to construct rich bidirectional models from text. Models are capable of learning information in both directions.

Tokenizing the input question, padding it to the maximum length, and adding a [CLS] token is the first stage in this procedure. An aggregated representation of the complete sequence is provided by the [CLS] token in BERT. In the next step, we use the tokenized sequence as an input for the BERT model. A neural network classifier is used for the [CLS] output of the BERT model to assign an input question to one of four preset categories in a given data set.

Visual and Textual Feature Extraction

Estimating suitable representations for questions and images is fundamental to the visual question answering (VQA) process to allow for easy processing by internal algorithms and deep neural architectures. Providing new ways of building such representations can be useful for processing both visual and textual data.

Given an image and a question in natural language, the VQA system tries to find the correct answer using the features of the image and data inferredfrom textual features. Feature extraction is a two-step process. First, we used EfficientNet as a deep learning model to extract visual features. For textual characteristics, we employed a bidirectional long short-term

memory (Bi-LSTM) model. An attention model was utilized to combine these two types of extracted characteristics.

EfficientNet for Visual Feature Extraction

Using a compound coefficient, EfficientNet scales depth, width, resolution in a CNN architecture. Compared to the usual practice, the EfficientNet scaling approach consistently scales these parameters using a given set of scaling coefficients. We may, for instance, raise the network depth by α, β, and γ if we wish to employ 2^N times more computing resources. These constant coefficients were obtained by conducting a short grid search on the original small model. Different scaling methods are shown in Figure 2.

It is logical to assume that if a picture is larger, the network requires more layers and channels in order to extend the receptive field and catch more fine-grained patterns. With the inclusion of squeeze-and-excitation blocks, the EfficientNet-B0 network was built on the inverted bottleneck residual blocks of MobileNetV2.

Bidirectional Long Short-Term Memory (BiLSTM) for Textual Feature Extraction

There are situations in which a traditional neural network may be used. RNNs are novel structures that emerge when experiments are conducted in the form of time sequences. As a result of the new structure, earlier inputs will be stored in the network's internal states, affecting all future outputs. This neural architecture comes in handy when dealing with "time series" data. LSTM neural networks are found in a variety of applications. According to this logic, it is crucial to know what happened beforehand in order to predict what will happen in the future. However, the hidden state of the LSTM allows it to retain information from previous inputs. Thus, LSTM is only able to store data from the past, because all of its inputs have been generated in the past.

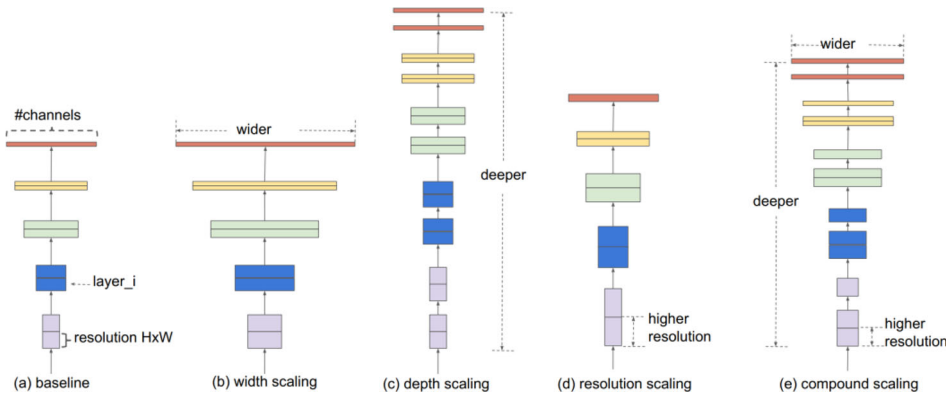

Figure 2. Different scaling methods vs. Compound scaling (Tan and Le 2019).

To resolve this problem, both future and previous inputs may be executed simultaneously with bidirectional usage. This means that inputs can be executed in two different directions. An LSTM that goes backwards keeps the knowledge it gained in the future, and, by merging the two hidden states, one can maintain the information from the past and the future at all times in the LSTM. Bidirectional RNNs are simply two separate RNNs combined. Networks can access past and future information at every time step because of this structure. This strategy differs from unidirectionality in that it preserves information from both the past and the future at any given moment in time, while unidirectionality merely preserves information from the past.

For question representation, we used a Bi-LSTM variant of RNNs, which captures long-range dependencies with their hidden states. Word embeddings are learned from word2vec and given as input to the Bi-LSTM. Word2vec is a prediction-based model designed to predict words based on their surrounding linguistic context by using one of two distinct neural network language models—skip-gram and Continuous Bag of Words—where, in each step, the neural network is trained with a set of words in the window (Altszyler, Sigman, and Slezak 2018). Our word2vec model used the skip-gram model with the default parameters. To learn the output vectors of skip-gram, we used the hierarchical softmax algorithm for better optimization.

Parameter Selection Based on a Genetic Algorithm

A genetic algorithm (GA) is a search method based on natural selection and genetics. In natural genetics, the presence or lack of specific genes and their chromosomal arrangement determine the unique characteristics of people in a given demographic. Different biological mechanisms affecting genetic structure pass on certain qualities from one generation to the next. The consequence is a well-adapted population, also known as "survival of the fittest."

This is also true of GA, which uses a finite length of string coding for each solution's parameters in the search problem. Each string represents a person, and the fitness value of each individual increases its power during the course of survival. Evolutionary performance improves with higher fitness values. A generation is made up of a certain number of people. Every generation, parents are chosen based on their fitness values, and strings of children are then generated via a variety of genetic operators. The new generation is created with the help of their computed fitness levels. As long as a certain threshold is reached, the process is repeated. When a population's average fitness increases over time, it reflects an improvement in overall quality of life.

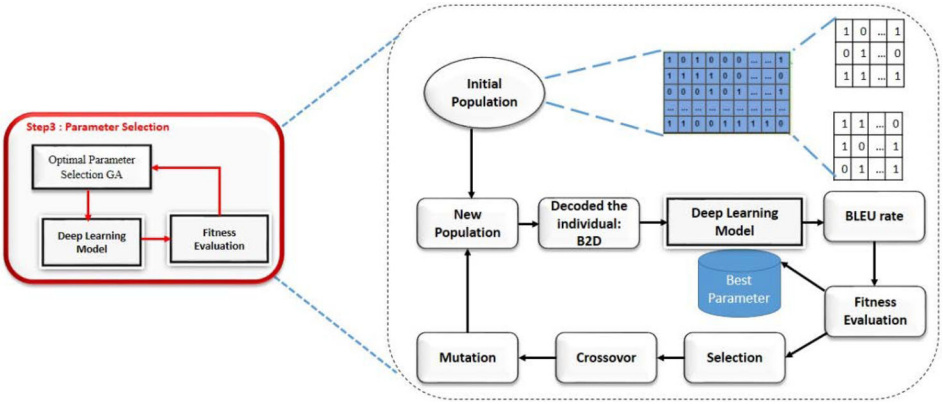

Figure 3. Parameter selection method based on an adaptive GA.

It is necessary to assess each individual's fitness in terms of the objective function in order to solve an optimization issue for each successive generation. Equation (1) defines the adaptive fitness function as a metric for assessing the quality of the solution represented by the GA.

$$\text{Fitness function} = W_1 * R_B + W_2 * 1/N_E \tag{1}$$

Where:

$$W_1 + W_2 = 1 \tag{1}$$

R_B represents the BLEU rate

N_E represents the epoch number

To apply the different steps of the GA presented in Figure 3, such as crossover and mutation, each individual within the overall population represents the number of epochs and the batch size coded in binary.

Experiment Results and Discussion

In this section, we present the results of the experiments performed on the model presented in this paper. We also compare the results obtained by several question classification methods. The proposed model was implemented in Python with an Rtx 2060 graphics card and 16 GB RAM.

To apply our proposed model, we carried out several series of experiments using deep learning models. Our model consists of three scenarios:

1. **Scenario 1**: Medical question classification based on the Bidirectional Encoder Representations from Transformers (BERT) model
2. **Scenario 2**: Visual question answering (VQA) based on optimal parameter selection method
3. **Scenario 3**: Comparing our best result with the official submissions to the ImageCLEF VQA task

Data Set Description and Evaluation Metrics

To assess our improved deep learning model, we used the ImageCLEF 2019 data set, which is available online. Table 1 provides a more detailed description of the data set. Each dataset's picture has a set of questions where each of one falls into one of the following four categories: plan, modality, organ and abnormality. In Table 2, we give an example of a sample picture and the four associated questions.

BiLingual Evaluation Understudy (BLEU) (Papineni et al. 2002), Word-Based Semantic Similarity (WBSS) (Sogancioglu, Öztürk, and Özgür 2017), and Accuracy were utilized to assess the suggested model's efficiency in VQA. A system-produced response was compared to the ground-truth answer using BLEU. WBSS (Wu and Palmer 1994) is the latest technique to compute semantic similarity in the biomedical sector based on the Wu & Palmer Similarity (WUPS) and the wordnet lexicon. Accuracy measurements for the generic VQA job take into account the precise matching of a supplied response and the ground-truth answer.

Table 1. ImageCLEF2019 Data set for VQA task (Tan and Le 2019).

ImageCLEF 2019 Data set			
	Training	Validation	Test
Images	3200	500	500
Questions	12792	2000	500

Table 2. Sample of Image and four question category from ImagCLEF Data Set.

Category	Plan	Modality	Organ	Abnormality
Question	what plane was used?	what type of imaging modality is shown?	what part of the body is being imaged?	what is abnormal in the ct scan?
Answer	axial	ct w/contrast (iv)	gastrointestinal	hepatic hemangioma

Medical Question Classification

Optimal selection parameters for deep learning models based on a genetic algorithm (GA) were the primary goal of this study; however, we began by classifying medical questions into four categories to increase the accuracy of our method. This was done by comparing a number of methods that fall within typical machine learning. Algorithms tested in this section include AdaBoost, K-Nearest Neighbors, and Naive Bayes. We employed typical measures, such as accuracy, precision, recall, and F-measure, to test our classification approach.

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

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

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

$$\text{ROC}/\text{AUC} = \text{TPR} = \text{TP}/(\text{TP} + \text{FN}) \tag{5}$$

Where $\text{TPR} = \text{FP}/(\text{FP} + \text{TN})$

Where TP stands for true positive, FP stands for false positive, P stands for precision, R stands for recall, TPR stands for true positive rate, and FPR stands for false positive rate.

Table 3 displays the results of the above-named algorithms. In terms of accuracy, the convolutional neural network (CNN)+BERT hybrid model had the highest score of 100, followed by the KNN approach, which received a score of **0.99**. The hard contextual relationship between the major sentences of the questions in each category may account for BERT's high score, especially given that BERT uses a transformer, an attention mechanism that learns the contextual connections between the words or subwords in a text.

Optimal Parameter Selection

The aim of this section is twofold: first, to compare different solutions found by the GA, such as optimal parameters; second, to evaluate the proposed optimal deep learning method for VQA. We evaluated our method by four medical question categories: plan, organ, modality, and abnormality. All results are presented in the below tables.

1. Table 4 for results obtained for questions from the plan category;
2. Table 5 for results obtained for questions from the organ category;
3. Table 6 for results obtained for questions from the modality category;
4. Table 7 for results obtained for questions from the abnormality category.

Table 3. Comparison of the proposed CNN-BERT model with other classifiers.

ImageCLEF2019 Data set				
Models	Accuracy	Precision	Recall	F1 Score
AdaBoost	75	62	75	66
K-Nearest Neighbors	99	99	99	99
Multinomial Naive Bayes	96	96	96	96
ImageCLEF	70	75	75	75
CNN + BERT	100	100	100	100

Table 4. Sample of optimal deep learning model results for plan question category.

Parameter selection (GA)		Question for plan category		
Epoch number	Batch size	Bleu	WBSS	Accuracy
150	256	0.52	0.52	0.52
60	256	0.52	0.52	0.52
150	32	0.62	0.62	0.62
60	32	**0.62**	**0.62**	**0.62**
60	64	0.61	0.61	0.61
60	16	0.60	0.60	0.62

Table 5. Sample of optimal deep learning model results for the organ category.

Parameter selection (GA)		Question for organ category		
Epoch number	Batch size	Bleu	WBSS	Accuracy
150	256	0.45	0.45	0.41
60	32	0.49	0.43	0.43
150	32	0.45	0.40	0.43
150	64	0.43	0.37	0.4
60	64	0.48	0.41	0.45
150	16	0.48	0.41	0.45

Table 6. Sample of optimal deep learning model results for the modality category.

Parameter selection (GA)		Question for modality category		
Epoch number	Batch size	Bleu	WBSS	Accuracy
150	256	0.28	0.32	0.23
150	32	0.48	0.45	0.34
150	**16**	**0.52**	**0.49**	**0.44**
60	16	0.43	0.42	0.35

Table 7. Sample of optimal deep learning model results for the abnormality category.

Parameter selection (GA)		Question for abnormality category		
Epoch number	Batch size	Bleu	WBSS	Accuracy
150	256	0.044	0.082	0.04
60	32	0.05	0.098	0.048
150	32	0.056	0.1	0.056
150	16	0.056	0.096	0.056
60	16	0.032	0.076	0.032
90	**32**	**0.063**	**0.11**	**0.056**

The results obtained by the optimal deep learning models based on an adaptive GA for VQA are presented in Tables 4–7. We noted that the best results were around **0.62**, and were obtained using BLEU in the plan

category. If we applied our approach based on the GA, we found that we could achieve a high rate for the different evaluation metrics with a minimal number of epochs. For the organ category, we reached a BLEU rate of around 0.49 for an epoch number of 60 instead of 150. This reduction in epoch number was also observed in both the modality and abnormality categories. For the organ category, we were able to reduce the number of epochs to 90, and with better accuracy.

It is clear how the results obtained for the three categories: plan, organ, and modality outperform the results obtained for the abnormality category. This can be explained by the high similarities between the medical images, especially from the abnormality category. The ImageCLEF data set comprises only 3,200 images; because of this, the similarities between abnormality images are very high. Despite the data augmentation method applied in our data set, we need a larger data set to effectively train our model. Though the results for the abnormality category are not sufficient, our model was able to improve the results through the selection of efficient parameters, increasing them from 0.043 to 0.063 by changing the batch size. Figure 4 represents the change in BLUE rate according to the epoch number and batch size selected by GA.

By optimizing the batch size, we can ensure the implementation of our model on a cheaper GPU, and, by optimizing the number of epochs, we also guarantee ourselves extra time. The results presented in the tables above confirm that our proposed optimal deep learning-based model ensures gains on two levels: GPU and time.

Due to the short quantity of data, the process of response generation is significantly more tied to the frequency of terms (such as "multiforme,"

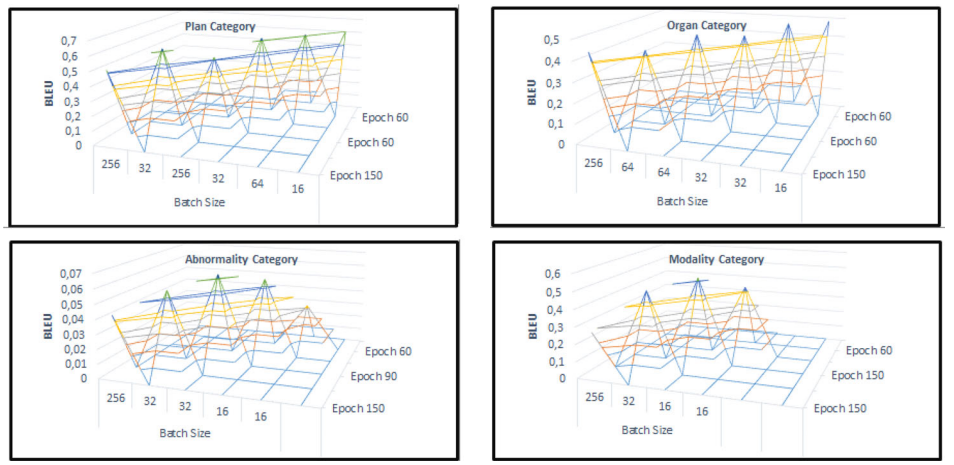

Figure 4. The blue rates obtained according to the selected optimal parameters: epoch number and batch size.

Table 8. Ranking of the proposal model run compared to official runs of ImageClef 2019 (Tan and Le 2019).

VQA-Med 2019: Accuracy scores				
	Modality	Plane	Organ	Abnormality
Proposal model	**0.52**	**0.62**	**0.48**	**0.06**
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

"glioblastoma," etc.) in the learning set than to medical pictures, which may explain the weak performance for the category "abnormality."

Comparison of the Proposal Model with Official Submissions of ImageClef 2019

To demonstrate the effectiveness of our proposed strategy, we compared its results with the runs of ImageCLEF 2019 participants. As illustrated in Table 8, ImageCLEF 2019s official relevance ratings put our run at the top of runs with the plan, organ, and modality categories. All the runs of ImageCLEF 2019 are implemented without using a classification process. This comparison proved that the classification of questions has a direct impact on the effectiveness of our model to predict the right answer. In fact, by classifying questions into categories our model is able to understand the intension behind users needs. Thus, as result, the answer search space will be reduced in a specific category for a given question (return the answer corresponding to that category only).

The effectiveness of our proposed model is also due to the GA's selection of the best parameters. This proves that two parameters have a significant impact on the results of the deep learning model. To further support this claim, we employed the BERT model's pre-step question-category classifier that has a powerful impact on the final outcome of the VQA model by achieving a high accuracy rate.

Conclusion

This paper presented a new parameter selection method for VQA deep learning model. Our method uses a GA to select the optimal epoch number and batch size. First, medical questions and images are mapped during pre-processing; then a question classification method is applied to assign each question to its appropriate category. The classification method is based on BERT model. We extract visual and textual features using EfficientNet and Bi-LSTM models, respectively. Finally, we combine all extracted features

and move them into a dense layer, as with a classification step, based on the softmax method. Before evaluating the proposed model for parameter selection, we performed a comparison between several classification methods to determine the best one. We then compared the parameters chosen by the GA during the parameter selection step. The results confirm that the use of a selection method improves the accuracy rate as well as saving time by minimizing the number of epochs.

In future works, several improvements can be made. We plan to address knowledge-based VQA. The information provided in the questions and the corresponding images are not always sufficient to predict the right answer, and answering the questions often requires external knowledge resources. For those reasons, we plan to deeply analyze the properties and relationships of the data set to extract related visual and semantic knowledge.

Acknowledgments

The authors extend their appreciation to the Deanship of Scientific Research at Jouf University for funding this work through research grant No (DSR-2021-02-0110).

References

Abacha, A. B., S. A. Hasan, V. Datla, J. Liu, D. Demner-Fushman, and H. MUller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019, in CLEF, 2019.

Agrawal, A., J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra. 2017. Vqa: Visual question answering. *International Journal of Computer Vision* 123 (1):4–31. doi: 10.1007/s11263-016-0966-6.

Al-Sadi, A., H. Al-Theiabat, and M. Al-Ayyoub. The inception team at vqa-med 2020: Pretrained vgg with data augmentation for medical vqa and vqg. CLEF, 2020.

Altszyler, E., M. Sigman, and D. F. Slezak. Corpus specificity in lsa and word2vec: The role of out-of-domain documents, in Rep4NLP@ACL, 2018.

Chen, G., H. Gong, and G. Li. Hcp-mic at vqa-med 2020: Effective visual representation for medical visual question answering, in CLEF, 2020.

Dalton, J., L. Dietz, and J. Allan. 2014. Entity query feature expansion using knowledge base links. Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, in NAACL, 2019.

Emilio, G.-G. Artificial intelligence in medicine and healthcare: applications, availability and societal impact, 2020.

Fukui, A., D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding, in EMNLP, 2016.

Gao, H., J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question, NIPS, 2015.

Ghourabi, A. 2021. A bert-based system for multi-topic labeling of arabic content. 2021 12th International Conference on Information and Communication Systems (ICICS), pp. 486–489, doi: 10.1109/ICICS52457.2021.9464540.

He, S., and D. Han. 2020. An effective dense co-attention networks for visual question answering. *Sensors* 20 (17):4897. doi: 10.3390/s20174897.

Ioannidou, A., E. Chatzilari, S. Nikolopoulos, and Y. Kompatsiaris. 2017. Deep learning advances in computer vision with 3d data. *ACM Computing Surveys* 50 (2):1–38. doi: 10.1145/3042064.

Ionescu, B., H. MUller, R. P'Eteri, A. B. Abacha, V. Datla, S. A. Hasan, D. DemnerFushman, S. Kozlovski, V. Liauchuk, Y. D. Cid, et al. 2020. Overview of the imageclef 2020: Multimedia retrieval in medical, lifelogging, nature, and internet applications, in CLEF.

Kafle, K., and C. Kanan. 2016. Answer-type prediction for visual question answering. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4976–4984. doi: 10.1109/CVPR.2016.538.

Kim, J.-H., J. Jun, and B.-T. Zhang. Bilinear attention networks, in NeurIPS, 2018.

Krishna, R., Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123 (1):32–73. doi: 10.1007/s11263-016-0981-7.

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, et al. 2015. Dbpedia – a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web* 6 (2):167–95. doi: 10.3233/SW-140134.

Li, R., and J. Jia. Visual question answering with question representation update (qru), in NIPS, 2016.

Liao, Z., Q. Wu, C. Shen, A. van den Hengel, and J. W. Verjans. Aiml at vqa-med 2020: Knowledge inference via a skeleton-based sentence mapping approach for medical domain visual question answering, CLEF, 2020.

Lu, J., J. Yang, D. Batra, and D. Parikh. Hierarchical co-attention for visual question answering, arXiv: Computer Vision and Pattern Recognition, 2016.

Ma, L., Z. Lu, and H. Li. Learning to answer questions from image using convolutional neural network, in AAAI, 2016.

Malinowski, M., M. Rohrbach, and M. Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1–9, doi: 10.1109/ICCV.2015.9.

Marino, K., M. Rastegari, A. Farhadi, and R. Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3190–3199, doi: 10.1109/CVPR.2019.00331.

Mohamed, S. S. N., and K. Srinivasan. Imageclef 2020: An approach for visual question answering using vgg-lstm for different datasets, in CLEF, 2020.

Nguyen, D.-K., and T. Okatani. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6087–6096,

Noh, H., P. H. Seo, and B. Han. 2016. Image question answering using convolutional neural network with dynamic parameter prediction. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 30–38. doi: 10.1109/CVPR.2016.11.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation, in ACL, 2002.

Rebele, T., F. M. Suchanek, J. Hoffart, J. A. Biega, E. Kuzey, and G. Weikum. Yago: A multilingual knowledge base from wikipedia, wordnet, and geonames, in SEMWEB, 2016.

Ren, M., R. Kiros, and R. S. Zemel. Image question answering: A visual semantic embedding model and a new dataset, ArXiv, vol. abs/1505.02074, 2015.
Richardson, M., and P. M. Domingos. 2006. Markov logic networks. *Machine Learning* 62 (1-2):107–36. doi: 10.1007/s10994-006-5833-1.
Saito, K., A. Shin, Y. Ushiku, and T. Harada. 2017. Dualnet: Domain-invariant network for visual question answering. 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 829–834. doi: 10.1109/ICME.2017.8019436.
Sarrouti, M. Nlm at vqa-med 2020: Visual question answering and generation in the medical domain, in CLEF, 2020.
Shen, K., and M. Kejriwal. A data-driven study of commonsense knowledge using the conceptnet knowledge base, ArXiv, vol. abs/2011.14084, 2020.
Shih, K. J., S. Singh, and D. Hoiem. 2016. Where to look: Focus regions for visual question answering. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4613–4621, doi: 10.1109/CVPR.2016.499.
Soganciogl u, G., H. Öztürk, and A. Özgür. 2017. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics* 33:i49–i58.
Tan, M., and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, ArXiv, vol. abs/1905.11946, 2019.
Teney, D., P. Anderson, X. He, and A. van den Hengel. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4223–32.
Torfi, A., R. A. Shirvani, Y. Keneshloo, N. Tavvaf, and E. A. Fox. Natural language processing advancements by deep learning: A survey. ArXiv, vol. abs/2003.01200. 2020.
Verma, H., and S. Ramachandran. Harendrakv at vqa-med 2020: Sequential vqa with attention for medical visual question answering, in CLEF, 2020.
Wang, P., Q. Wu, C. Shen, A. R. Dick, and A. van den Hengel. Explicit knowledge-based reasoning for visual question answering, in IJCAI, 2017.
Wang, P., Q. Wu, C. Shen, A. R. Dick, and A. van den Hengel. 2018. Fvqa: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (10):2413–27. doi: 10.1109/TPAMI.2017.2754246.
Wu, Q., D. Teney, P. Wang, C. Shen, A. R. Dick, and A. van den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding* 163:21–40. doi: 10.1016/j.cviu.2017.05.001.
Wu, Q., P. Wang, C. Shen, A. R. Dick, and A. van den Hengel. 2016. Ask me anything: Freeform visual question answering based on knowledge from external sources. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4622–4630, doi: 10.1109/CVPR.2016.500.
Wu, Z., and M. Palmer. Verb semantics and lexical selection, ArXiv, vol. abs/cmplg/9406033, 1994.
Yu, Z., J. Yu, Y. Cui, D. Tao, and Q. Tian. 2019. Deep modular co-attention networks for visual question answering. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6274–6283, doi: 10.1109/CVPR.2019.00644.
Zhou, B., Y. Tian, S. Sukhbaatar, A. D. Szlam, and R. Fergus. Simple baseline for visual question answering, ArXiv, vol. abs/1512.02167, 2015.
Zhu, Y., C. Zhang, C. R'e, and L. Fei-Fei. Building a large-scale multimodal knowledge base system for answering visual queries, arXiv: Computer Vision and Pattern Recognition, 2015.