Article

Goal-driven visual question generation from radiology images

Mourad Sarrouti [1,‡]*, Asma Ben Abacha [1,‡]* and Dina Demner-Fushman [1,‡]

[1] U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD, USA; sarrouti.mourad@gmail.com, asma.benabacha@gmail.com, ddemner@mail.nih.gov

* Correspondence: sarrouti.mourad@gmail.com; 301 518 1033 (M. S.), asma.benabacha@gmail.com (A. BA.)

‡ These authors contributed equally to this work.

Abstract: Visual Question Generation (VQG) from images is a rising research topic in both fields of natural language processing and computer vision. Although there are some recent efforts towards generating questions from images in the open domain, the VQG task in the medical domain has not been well-studied so far due to the lack of labeled data. In this paper, we introduce a goal-driven VQG approach for radiology images called VQGRaD that generates questions targeting specific image aspects such as modality and abnormality. In particular, we study generating natural language questions based on the visual content of the image and on additional information such as the image caption and the question category. VQGRaD encodes the dense vectors of different inputs into two latent spaces, which allows generating, for a specific question category, relevant questions about the images, with or without their captions. We also explore the impact of domain knowledge incorporation (e.g., medical entities and semantic types) and data augmentation techniques on visual question generation in the medical domain. Experiments performed on the VQA-RAD dataset of clinical visual questions showed that VQGRaD achieves 61.86% BLEU score and outperforms strong baselines. We also performed a blinded human evaluation of the grammaticality, fluency, and relevance of the generated questions. The human evaluation demonstrated the better quality of VQGRaD outputs and showed that incorporating medical entities improves the quality of the generated questions. Using the test data and evaluation process of the ImageCLEF 2020 VQA-Med challenge, we found that relying on the proposed data augmentation technique to generate new training samples by applying different kinds of transformations, can mitigate the lack of data, avoid overfitting, and bring a substantial improvement in medical VQG.

Keywords: Visual Question Generation; Visual Question Answering; Variational Autoencoders; Radiology Images; Domain Knowledge; UMLS; Data Augmentation; Computer Vision; Natural Language Processing; Artificial Intelligence; Medical Domain.

1. Introduction

Recent advancements in computer vision [1–3], natural language processing [4–6] and deep learning [7,8] research have enabled enormous progress in many medical image interpretation technologies that support clinical decision making and improve patient engagement [9–12].

Generating natural language questions for image understanding is a rising research topic in both the fields of natural language processing and computer vision [13,14]. The task, known as Visual Question Generation (VQG), has two main motivations. First, it supports creating large-scale collections of Visual Question Answering (VQA) pairs at low cost since VQG could automatically generate questions about an image. Second, it can also play a role in improving the efficiency of human annotation for VQA datasets construction [15]. VQG combines natural language processing that provides the ability to generate the question, and computer vision techniques that allow the understanding of the image's content.

In contrast to answering visual questions about images, generating questions has received little attention so far. A few recent works have attempted to generate questions

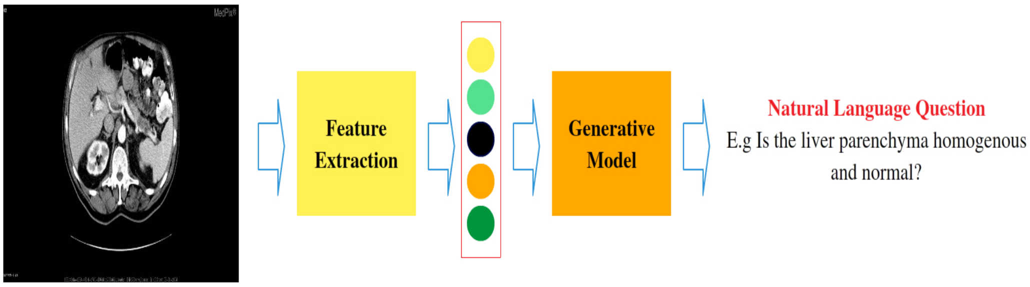

Feature
Extraction
Generative
Model
Natural Language Question
E.g Is the liver parenchyma homogenous and normal?

Figure 1. The VQG pipeline: image is taken as input to generate the question.

from images in the open domain [16–18]. However, the task of VQG in the medical domain has not been well-studied. In addition to the two main motivations mentioned above, VQG could benefit both doctors and patients. For example, patients could use the questions provided by VQG systems to start a conversation with their doctors and understand better their medical images. Moreover, such VQG systems could support medical education and clinical decision making by understanding medical images and generating questions related to their content [19].

The VQG task, as shown in Figure 1, consists of three main phases: (1) generating a representations of the image; (2) producing the embeddings by a neural network, and then (3) generating the question.

One major problem with medical VQG is the lack of large-scale labeled training data, which usually requires huge efforts to build, especially in the medical domain where domain experts are needed for data construction. Although deep learning models have achieved a remarkable success in computer vision and natural language processing tasks, the performance often depends on the size and quality of available training data, which is often tedious to collect [9,20,21]. Usually, to avoid the overfitting problem, the neural networks have to access more training data. But, many tasks lack access to large amounts of data, such as medical VQG and VQA.

Recently, we have presented a VQG system that is able to generate questions when shown radiology images [22]. However, this approach is not goal-driven as it does not guarantee that the generated questions will address a specific aspect of the image. Our previous approach tackled generating natural language questions that are relevant to radiology images without any constraints on the types of the generated questions. Developing a method capable of asking goal-oriented questions about images is a challenging research problem. Towards this end, this work aims to develop an approach to generating questions that ask specific information about radiology images. As shown in Figure 2, by specifying the type of the expected questions, different questions can be generated for a given image such as "What is the condition seen in this image" for abnormality, "Does this have contrast" for modality. The goal-driven question generation process allows, for a given image, to specify the category of the expected question. Such an approach will allow better control of the generated questions from radiology images when they involve multiple topics of interest, which occurs often in medical images.

In this paper, we introduce VQGRaD, a goal-driven VQG system for generating natural language questions about radiology images. VQGRaD is tasked with generating a natural language question when provided with an image (with or without its caption) and the question category. In summary, this paper makes the following contributions:

1. To the best of our knowledge, this work is the first attempt to generate visual questions about medical images that will result in a specific type of answer when provided with an initial indication of the question category/type.
2. To overcome the data limitation of VQG in closed domains, we propose a new data augmentation method for natural language questions.

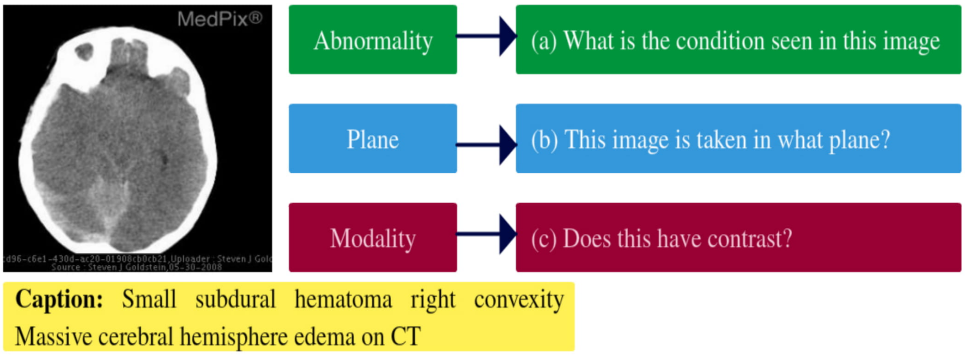

MedPix®
Abnormality
(a) What is the condition seen in this image
Plane
(b) This image is taken in what plane?
Modality
(c) Does this have contrast?
Caption: Small subdural hematoma right convexity Massive cerebral hemisphere edema on CT

Figure 2. Examples of goal-driven questions about a radiology image. The possible questions should be relevant to the given category, also known as question type. By specifying the type of the expected question, different questions such as (a), (b) and (c) can be generated from the same radiology image.

3. VQGRaD is designed to work with or without an image caption and only requires the image and the question category as minimal inputs.
4. We study the impact of domain knowledge incorporation, such as named entities and semantic types, in the proposed VQGRaD approach.
5. VQGRaD is evaluated using the VQA-RAD dataset of clinical questions and radiology images. Experimental results show that VQGRaD performs better than strong baselines, with a BLEU-1 score of 61.86%. We also report the VQG models' results in our participation at the VQA-Med 2020 challenge.
6. We perform a manual evaluation to study the grammaticality, fluency, and relevance of the generated question from radiology images.

The remainder of the paper is organized as follows. First, related work concerning the main visual question generation systems is reviewed in Section 2. Then, the proposed methods are presented in Section 3. Several comprehensive experiments are performed to evaluate the effectiveness of the proposed methods in Section 4 where experimental settings, evaluation metrics, benchmark datasets, and results are presented. Conclusion and future work are finally presented in Section 5.

2. Related work

Question generation (QG) is the task of automatically creating natural language questions from a range of inputs, such as natural language text [23–25], structured data [26] and images [13,16]. While many natural language processing and computer vision problems involve extracting information from the texts and images such as VQA [11,27,28], VQG, which can be considered as a complementary task of VQA, is a multi-modal problem involving image understanding and natural language generation, especially using generative methods.

VQG in the open-domain benefited from the available large-scale annotated datasets [29–31]. These large-scale datasets allow for a variety of work studying generative models and continuous latent spaces for generating visual questions in the open domain [14,32]. Early work on goal-oriented visual question generation focused primarily on reinforcement learning setting [33,34]. Recent VQG approaches have used autoencoders architecture to generate questions from images and some additional inputs such as answers and categories of questions [15,16,35]. The successes of these systems have primarily been a result of variational autoencoders [36].

In this work, we are interested in generating questions in closed domains. Visual question generation in closed domains, such as the medical field, is a challenging task [37–40] that is still understudied. For instance, Lau *et al.* [19] created the first VQG dataset in the medical domain (VQA-RAD), where each radiology image was manually annotated with several questions. However, this dataset is too small for training efficient VQG models. Recently, we have developed a VQG system that is able to generate questions from radiology images [22]. However, this approach is not goal-driven as it does not guarantee that the generated question will address a specific aspect of the image.

Inspired by the aforementioned open-domain research, we present in this paper VQGRaD, a goal-driven visual question generation system for radiology images based on the variational autoencoders architecture. Our work extends our previous method [22] by tackling visual question generation as a process that considers the question's category, the image and its caption.

VQGRaD is able to generate questions about four main categories: abnormality, modality, plane, and organ. These are the most frequent question categories in the VQA-RAD dataset. The questions can be generated from either (i) the image and the question category or (ii) the image, its caption, and the question category.

In addition, to overcome the data limitation problem in the medical domain, we propose a text-based augmentation method to automatically create new training questions. Data augmentation, the application of one or more deformations to labeled data which result in new, additional training data, is a promising solution to handle the data insufficiency problem [41,42].

Automatic data augmentation based on images is commonly used in computer vision [9,20,21] and can help train deep learning models, particularly when using smaller datasets. Simply flipping or shifting images can help the models to better learn by increasing the number of training images. However, the lack of available training images for VQG in the medical domain makes image-based data augmentation alone insufficient for boosting performance on the visual question generation task. Thus, training our supervised models on the augmented natural language data can allow them to become more invariant to these deformations and generalize better to unseen data.

Our text data augmentation method can also be used in open-domain and restricted-domain NLP tasks, such as text classification and question answering, as it relies on general morpho-syntactic features to replace relevant target words in the original text with words that have a high contextual similarity. The following section will present our proposed methods in detail.

3. Methods

The first goal of this study is to generate natural language questions that ask about specific topics such as modality, abnormality, plane, and organ.

To address this challenge, we present VQGRaD, a goal-driven visual question generation system for radiology images that aims to generate relevant questions based on the visual content of the image.

3.1. Problem Modeling

Given a pair (C, I), where C is the question category accompanied by a medical image I, VQGRaD is tasked with generating the appropriate question Q that will result in a specific type of answer. Mathematically, the VQG task can be formulated as:

$$Q = f(C, I, \alpha) \tag{1}$$

where f is the question generation function and α denotes the parameters of the model. Categories in C are modality, abnormality, plane, and organ.

In the following sections, we will provide a detailed description of our proposed methods.

3.2. VQGRaD

The VQGRaD model is based on the variational autoencoders architecture [36]. It first encodes the image and its caption along with the category before generating

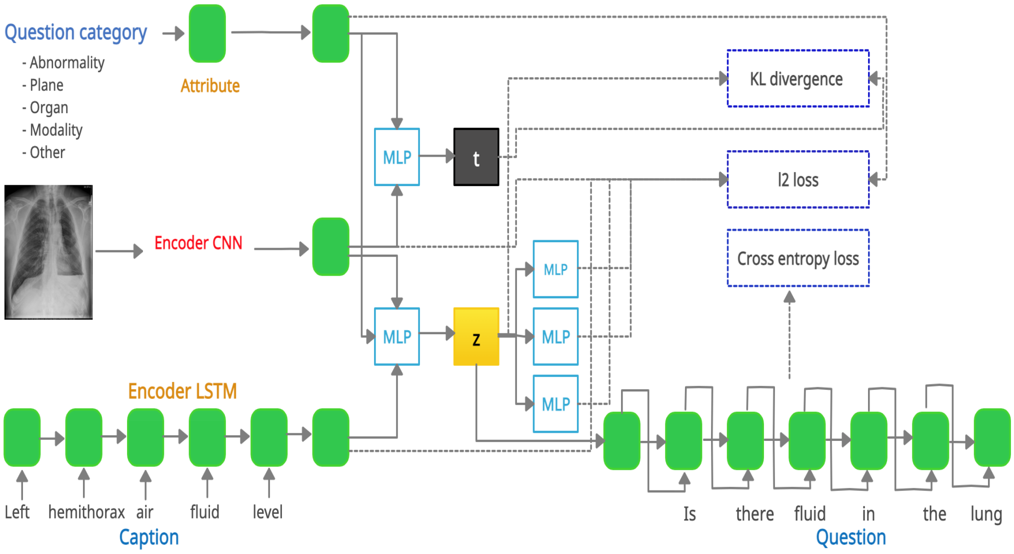


Figure 3. Overview of VQGRaD, a visual question generation system for radiology images. Input captions are encoded by an LSTM. Images are encoded by a CNN. t space contains images and categories features, whereas z space includes images, captions, and categories features. Questions can be generated from either the caption latent space z or the category latent space t.

the question. VAEs comprise two neural network components, known as encoder and decoder, for learning the probability distributions of data $p(x)$. The encoder transforms the latent variable z that is created from raw data x into latent space $z - space$. In other words, the encoder compresses the data from the initial space to the encoded space, also called latent space. The decoder, on the other hand, aims at recovering x using z extracted from the latent space. The training of the encoder and decoder proceeds by maximizing marginal likelihood $\log p(x)$. Finding the Evidence Lower BOund (ELBO) yields:

$$\begin{aligned}\log p(x) &\geq E_{z \sim q_\theta(z|x)}[\log p_\phi(x|z)] - KL(q_\theta(z|x)||p(z)) \\ &= ELBO\end{aligned} \tag{2}$$

Where $q(z|x)$ and $p(x|z)$ are the probability distributions of the encoder and the decoder, respectively.

The loss function that is minimized when training VAEs is the negative log-likelihood with a regularizer. It consists of a reconstruction loss and a regularisation loss (on the latent layer). Reconstruction loss consists at making the scheme for encoding and decoding as performant as possible, whereas the regularisation loss regularises the latent space organisation by making the distributions returned by the encoder close to a standard normal distribution. The loss function l_i for datapoint x_i is:

$$l_i(\phi, \theta) = -E_{z \sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)] + KL(q_\theta(z|x_i)||p(z)) \tag{3}$$

where $E_{z \sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)]$ is the reconstruction error and $KL(q_\theta(z|x)||p(z))$ is the Kullback-Leibler divergence regularization between the returned distribution and a standard Gaussian. ϕ and θ, the parameters for the decoder distribution $p_\phi(x|z)$ and the encoder distribution $q_\theta(z|x)$ respectively.

In VQGRaD, as shown in Figure 3, a Convolutional Neural Network (CNN) is used to obtain the image feature map v and a Long Short Term Memory network (LSTM) [43] is used to generate the embedded caption features c. The categories of the questions are represented as a one hot vector a. It then encodes the dense vectors h_c, h_a and h_v of the caption, the category, and the image, respectively, into a continuous, dense, latent z-space. It also encodes the dense vectors h_a and h_v into another continuous, dense, latent t-space based on the continuous latent space introduced in [16] for regularization. This allows our system to maximize the mutual information $MI(.)$ between the encoded features, i.e., the image, the caption, the category, and the latent space. $MI(.)$ measures how much knowing one of the predefined features reduces uncertainty about the other. For example, if h_a and h_v are independent, then knowing h_v does not give any information about h_a and vice versa, so their mutual information is zero.

In our case, the optimization is computed as follow:

$$\begin{aligned} &\max_{\phi} MI(q,z|c,a,v) + \lambda_1 MI(c,z) + \lambda_2 MI(a,z) + \lambda_3 MI(v,z) \\ &s.t. |z| = p_\phi(q|z) \end{aligned} \tag{4}$$

where λ_1, λ_2, λ_3 are hyperparameters that relatively weight $MI(.)$ terms in the optimization. $p_\phi(q|z)$ is the learned mapping, parameterized by ϕ, from the image, the caption, and the category to this latent space.

Then, VQGRaD reconstructs the inputs from the z-space using a simple Multi Layer Perceptron (MLP) which is a neural network with fully connected layers. It generates the reconstructed image, caption, and category features L_v, L_c, L_a, and optimizes the model by minimizing the following l_2 losses:

$$\begin{aligned} L_v &= ||h_v - \hat{h_v}||_2 \\ L_c &= ||h_c - \hat{h_c}||_2 \\ L_a &= ||h_a - \hat{h_a}||_2 \end{aligned} \tag{5}$$

On the other hand, VQGRaD trained $t-$space by minimizing the KL-divergence with $z-$space:

$$\begin{aligned} L_t &= KL(p_\phi(z|c,a,v), p_\theta(t|a,v)) \\ &= \log\sigma_t - \log\sigma_p + \frac{\sigma_z + (\mu_t - \mu_z)}{2\sigma_t} - 0.5 \end{aligned} \tag{6}$$

where ϕ and θ are the parameters used to embed into $z-$space and $t-$space, respectively. We used the reparameterization trick [36], to generate means μ_z and standard deviations σ_z, combine it with a sampled unit Gaussian noise ϵ to generate:

$$z = \mu_z + \epsilon\sigma_z \tag{7}$$

In VQGRaD, the $t-$space is not only used for regularization but also to generate questions from only the image and the question category.

Finally, VQGRaD uses an LSTM decoder to generate the question $\hat{q}$ from either the z-space or the $t-$space. The decoder takes a sample from the latent dimension z-space, and uses that as an input to output the question $\hat{q}$. It receives a "start" symbol and proceeds to output a question word by word until it produces an "end" symbol. We used Cross Entropy loss function to evaluate the neural network's quality and minimize the error L_g between the generated question $\hat{q}$ and the ground truth question q. The generation of each word of the question can be written as:

$$\hat{w}_t = \arg\max_{w \in \mathbb{W}} p(w|v, w_0, ..., w_{t-1}) \tag{8}$$

where $\hat{w}_t$ is the predicted word at t step, $\mathbb{W}$ denotes the word vocabulary, and $\hat{w}_i$ represents the i-th ground-truth word.

The final loss of VQGRaD is as follows:

$$L_{vqgrad} = \lambda_5 L_g + \lambda_4 KL + \lambda_3 L_v + \lambda_2 L_a + \lambda_1 L_c + \lambda_6 L_t \tag{9}$$

where KL is Kullback-Leibler divergence, $\lambda_1, \lambda_2, \lambda_3$ have already been introduced and $\lambda_4, \lambda_5, \lambda_6$ are hyperparameters that control the variational loss, the question generation loss, and the amount of regularization used in our model, respectively.

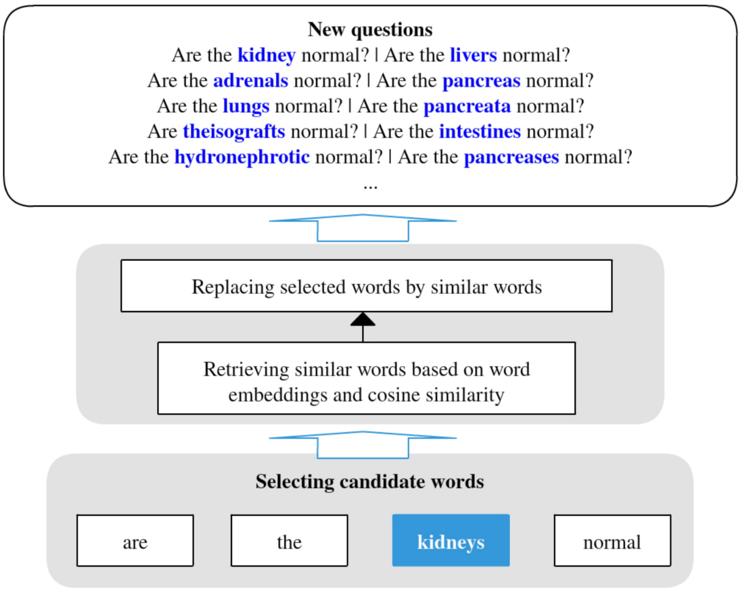

New questions
Are the kidney normal? | Are the livers normal?
Are the adrenals normal? | Are the pancreas normal?
Are the lungs normal? | Are the pancreata normal?
Are theisografts normal? | Are the intestines normal?
Are the hydronephrotic normal? | Are the pancreases normal?
...
Replacing selected words by similar words
Retrieving similar words based on word embeddings and cosine similarity
Selecting candidate words
are
the
kidneys
normal

Figure 4. Contextual augmentation, when a clinical question "Are the kidneys normal?" is augmented by replacing only selected words with similar words retrieved based on the cosine similarity and pretrained word embeddings.

3.3. Data Augmentation

Questions. For a given medical question q, we generate a set of new questions. During the augmenting process, we use the whole training data $D = \{q_i\}_{i=1}^{n}$ where n is the number of training questions. We expand each training question q_i into a set of instances q_i^k where k is the number of derived pairs for each training question. To do so, we first select nouns and verbs as candidate words, using the following part-of-speech tags[1]:

- NN: Noun, singular or mass.
- NNS: Noun, plural.
- NNP: Proper noun, singular.
- NNPS: Proper noun, plural.
- VBD: Verb, past tense.
- VBP: Verb, non-3rd person singular present.
- VBN: Verb, past participle.
- VBG: Verb, gerund or present participle.
- VBZ: Verb, 3rd person singular present.
- VB: Verb, base form.

Each candidate word is then replaced by contextually similar words using Wiki-PubMed-PMC embedding which was trained using four million English Wikipedia, PubMed, and PMC articles. Similar words for a given word are retrieved from the word embeddings space using cosine similarity. We compute the cosine similarity between a weight vector of the given word w_i in the question and the vectors for each word w_j in the pre-trained word embeddings. We use the top k similar words according to the cosine similarity. Several experiments were carried out with $k = \{5, 10, 15, 20, 30\}$ and found that the best result can be achieved with $k = 10$. Figure 4 presents some examples of created questions for the input question "Are the kidneys normal?".

Images. We also generate new training instances based on image augmentation techniques. To do so, we apply flipping, rotation, shifting, blurring techniques on the whole VQA-RAD training images. Figure 5 presents some examples of created images.

[1] We used NLTK [44] to perform part-of-speech tagging

doi:10.20944/preprints202107.0385.v1

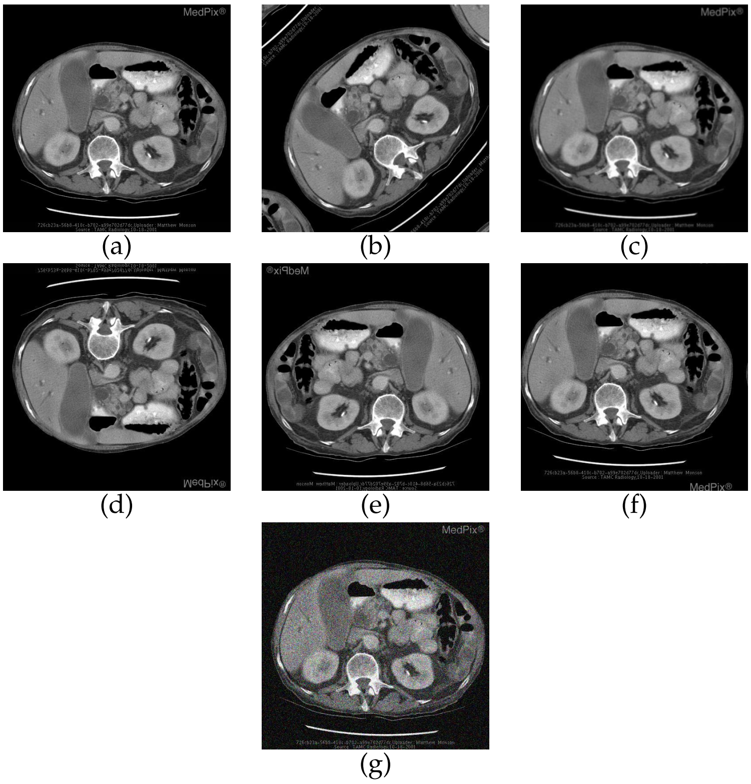


Figure 5. Examples of created images from the original image (a): (b) is the rotated image, (c) the blurred image, (d) the horizontally flipped image , (e) the vertically flipped image, (f) the shifted image, and (g) the noisy image.

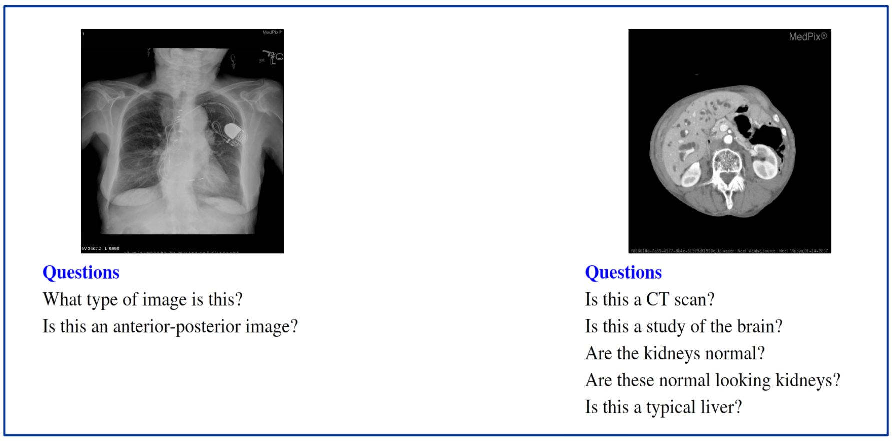
4. Experimental Settings and Results

In this section, we present our VQG results and conduct a comprehensive ablation analysis. As mentioned above, the proposed method is evaluated on the VQA-RAD and VQA-Med 2020 datasets.

4.1. Datasets

In this study, we used the VQA-RAD [19] dataset of clinical visual questions to evaluate our VQG system. The dataset contains 315 images and 3,515 corresponding questions. Figure 6 presents simple images and questions. Each image is associated with more than one question, each of which is accompanied with its category. In this work, we are particularly interested in five categories of questions: 'Modality", "Abnormality", "Organ", "Plane" and "Other". Table 1 presents the number of questions and images associated to each of these categories before and after data augmentation. The test set contains 100 reference questions with associated categories and images.

We have also used the datasets provided by the VQA-Med 2020 challenge at ImageCLEF 2020 during our participation. Given a radiology image, the VQG task consists of generating a natural language question based on the image's content. The dataset used in VQA-Med 2020 consists of 780 radiology images with 2,156 associated questions as training data, 141 radiology images with 164 questions as validation data, and 80 radiology images as test data. There are 1,942 unique questions in the 2,156 training questions. Some questions are associated with more than one image (up to 8 images). After applying data augmentation, our final training set consists of 161,348 questions. Figure 7 shows examples from VQA-Med 2020 VQG data.

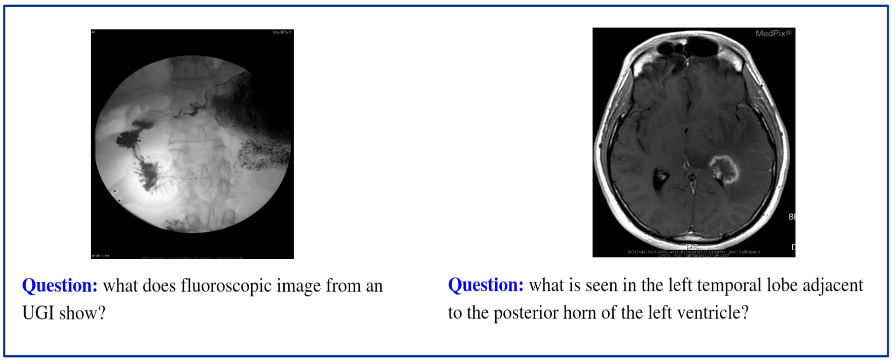

Questions
What type of image is this?
Is this an anterior-posterior image?

Questions
Is this a CT scan?
Is this a study of the brain?
Are the kidneys normal?
Are these normal looking kidneys?
Is this a typical liver?

Figure 6. Sample radiology images and the associated questions from the VQA-RAD dataset.

Table 1: The number of questions and images associated with each category. The values after "/" represent the number of questions and images created by our data augmentation techniques.

Category	#Questions	#Images
Abnormality	397/18642	112/784
Modality	288/5534	54/378
Organ	73/16408	135/945
Plane	163/9216	99/693
Other	348/19798	81/567
Total	1269/69598	239/1673


Question: what does fluoroscopic image from an UGI show?

Question: what is seen in the left temporal lobe adjacent to the posterior horn of the left ventricle?

Figure 7. Example of radiology images and the associated questions from the VQG training set of ImageCLEF 2020 VQA-Med.

4.2. Evaluation metrics

To investigate the performance of our visual question generation model, we make use of both automatic and manual evaluations.

4.2.1. Automatic evaluation

VQG is a sequence generation problem. Therefore, in the automatic evaluation, we used various language modeling evaluation metrics such as BLEU, ROUGE, METEOR, and CIDEr to measure the similarity of the system-generated questions and the ground truth questions in the test set. We used the evaluation package published by [45]. BLEU-{1-4} measures the quality of the generated question by counting the matching {1-4}-grams in the generated question to the {1-4}-grams in the reference question, respectively. METEOR compares the generated question with the reference question in terms of exact, stem, synonym, and paraphrase matches between words and phrases. ROUGE-L assesses the generated question based on the longest common subsequence shared by both the candidate and the reference question. The CIDEr measures consensus in questions by performing a Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n-gram.

4.2.2. Human evaluation

We also performed a human evaluation to measure the quality of the questions generated by our system and the baseline. To do so, we followed the standard approach in evaluating text generation systems [46], as used for question generation by [47,48]. We manually checked the generated questions and rated them in terms of relevancy, grammaticality, and fluency. The relevancy of a question is determined by the relationship between the question, the image, and the category. Grammaticality refers to the conformity of a question to the grammar rules. Fluency and common sense (readability) refers to the way individual words sound together within a question. Two experts at the U.S National Institutes of Health (NIH) performed manual evaluation. For each measure, the assessors were required to give a rating ranging from 1 to 3 scale (1 = Incorrect, 2 = Average (minor errors), 3 = Correct).

4.3. Implementation details

VQGRaD. Our VQGRaD is implemented using PyTorch. We used ImageNet-pretrained ResNet-50 [49] without fine-tuning its weights. Since the model expects an input of dimension $224 * 224$, we resized the input images to suit that dimension. The z-space and t-space are 100 dimensions, the Adam optimiser [50] with a learning rate of 0.0001, a batch size of 32, maximum sequence length for outputs of 20 tokens were used. All models were trained for 40 epochs using single P100 GPUs (16 GB VRAM) on a shared cluster, and the best results were used as final results. We optimized the hyperparameters such that $\lambda_1 = 0.001$, $\lambda_2 = 0.005$, $\lambda_3 = 0.001$, $\lambda_4 = 0.0001$ $\lambda_5 = 0.001$ and $\lambda_6 = 0.001$ for a total of 20 epochs. The source code are publicly available on GitHub at https://github.com/sarrouti/vqgrad (**the source code will be available upon acceptance of the paper**).

VQG baseline. We used our recent VQG system named VQGR [22] as a baseline. This model is based on the variational autoencoder architecture that takes an image as input and generates a question. In our implementation, we used ImageNet-pretrained ResNet-50 [49] provided by PyTorch without fine-tuning its weights as the image encoder and an LSTM decoder for generating questions. The source code is publicly available on GitHub at https://github.com/sarrouti/vqgr.

4.4. Experiments and Results

In order to study the task of visual question generation about radiology images and explore the impact of domain knowledge incorporation such as medical entities and

UMLS semantic types, we perform several experiments with different settings as shown in Table 2:

- **VQGRaD** is our full model that can generate questions from either the caption latent space z (image, caption, and category) or the category latent space t (image and category).
- **VQGRaD**$_{w\_t}$ includes another LSTM encoder to encode the image titles.
- **VQGRaD**$_{w\_st}$ includes another LSTM encoder to encode the UMLS semantic types extracted from the image captions.
- **VQGRaD**$_{w\_e}$ uses only UMLS entities instead of using all words in captions. PyMetamap[2], a python wrapper for MetaMap [51], has been used for extracting UMLS entities and semantic types.
- In **VQGRaD**$_{cap\_or\_c}$, the z-space contains only image and caption features.

All systems can generate questions from either the caption latent space z or the category latent space t.

Table 2 also presents a comparison of our proposed models and the baseline systems:

- The **VQGR** baseline system is trained on the VQA-RAD dataset without data augmentation.
- **VQGR**$_{w\_im\_aug}$ is trained on the dataset generated by augmenting the images.
- **VQGR**$_{w\_our\_aug}$ is trained on the dataset generated by our data augmentation technique.

By comparing VQGR, VQGR$_{w\_im\_aug}$ and VQGR$_{w\_our\_aug}$, we can see that our data augmentation technique helped considerably producing a significant improvement in the results. The best BLEU-1 score, 55.05%, was achieved using our data augmentation technique.

Furthermore, it is interesting to see that VQGRaD performs the best over the baseline systems and on all evaluation metrics. Moreover, all of our VQGRaD models outperform the baseline system by a significant margin. This confirms our hypothesis that the task of visual question generation can be goal-driven. VQGRaD achieved consistently better scores among other ablations when the questions were generated from the t-space, which contains the image and the question category features.

When adding the UMLS semantic types extracted from the captions in VQGRaD$_{w\_st}$ or the image titles in VQGRaD$_{w\_t}$, the models' performance was continuously improved in most metrics when the questions were generated from the $z-$space *(caption latent space)*. This is likely because the questions were generated from a latent space that encodes more features (including images, question category, caption or title, and UMLS semantic types) than in the VGGRaD system. However, the best results were obtained by VGGRaD when the questions were generated from the t-space *(category latent space)*. Thus, building end-to-end VQG models that consider the question type is a feasible and efficient task.

For additional evaluation, we used our VQG system during our participation in the VQG task of the VQA-Med challenge at ImageCLEF 2020 [52]. Given a radiology image, the VQG task consists of generating a natural language question based on the image's content. As the questions were all about abnormality, we only used our recent VQGR system based on VAEs and data augmentation, which takes an image as input and generates a natural language question as output. Table 3 shows our official results on the validation set at the VQG task of the VQA-Med challenge.

The official results of ImageCLEF 2020 VQA-Med showed that using a sequence generation model to solve VQG in the medical domain is complicated due to the problem of labeled data scarcity. Hence, the participating systems have used image classification approaches [53] to solve the VQG task. Small datasets might require models that have low complexity. Whereas sequence generation models require a large amount of training

[2] https://github.com/AnthonyMRios/pymetamap

Table 2: Ablation study and comparison of VQGR (baseline) and VQGRaD systems

	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
	VQGR	31.45	14.60	7.82	3.27	10.43	38.80	21.19
	VQGR$_{w\_im\_aug}$	44.83	30.10	23.62	19.81	18.98	24.30	23.43
	VQGR$_{w\_our\_aug}$	55.05	43.39	37.98	34.54	29.35	56.34	31.18
t-space	VQGRaD$_{cap\_or\_c}$	58.69	48.82	44.08	40.92	31.71	59.70	35.74
	VQGRaD$_{w\_t}$	57.12	46.62	41.49	37.96	29.61	58.93	34.99
	VQGRaD$_{w\_st}$	56.86	45.73	40.58	37.25	29.99	58.42	36.21
	VQGRaD$_{w\_e}$	61.81	**51.69**	46.69	43.35	**33.94**	63.62	40.33
	VQGRaD	**61.86**	51.65	**46.70**	43.40	33.88	**63.75**	**41.13**
z-space	VQGRaD$_{cap\_or\_c}$	59.31	49.31	44.73	41.75	32.54	60.90	36.38
	VQGRaD$_{w\_t}$	58.49	48.26	43.87	41.21	31.79	58.81	36.03
	VQGRaD$_{w\_st}$	60.74	50.06	45.00	41.90	32.81	61.36	36.89
	VQGRaD$_{w\_e}$	60.52	50.61	46.31	**43.68**	33.07	61.29	37.86
	VQGRaD	59.11	47.44	41.78	37.85	31.00	60.39	36.79

Table 4: Results of the manual evaluation of the best VQGRaD models and the VQG baseline system. "Relevancy", "Fluency", and "Grammaticality" are rated on a 1–3 scale (3 for the best). "Score" is the average of relevancy, fluency, and grammaticality scores. All numbers are normalized (divided by 60). The perfect score is 100.

Model	Relevancy	Grammaticality	Fluency	Score
VQGR	78.3	93.3	80.0	83.3
VQGRaD$_{w\_st(z-space)}$	83.3	92.5	91.6	89.16
VQGRaD$_{w\_e(z-space)}$	81.6	92.5	93.3	89.16
VQGRaD$_{w\_e(t-space)}$	96.6	96.6	97.5	96.9
VQGRaD$_{(t-space)}$	86.6	96.6	92.5	91.9

Table 5: Inter-rater reliability. We used F1-score to compute the inter-annotator agreement [54].

Model	Relevancy	Grammaticality	Fluency
VQG	0.42	0.27	0.51
VQGRaD$_{w\_st(z-space)}$	0.32	0.64	0.40
VQGRaD$_{w\_e(z-space)}$	0.40	0.64	0.48
VQGRaD$_{w\_e(t-space)}$	0.32	0.72	0.33
VQGRaD$_{(t-space)}$	0.35	0.72	0.43

data as they try to deeply learn the underlying data distribution of the input to output new sequences. The available training data for VQG in the medical domain is not large/varied enough for training a seq2seq model. However, once we increased the size/variance in the dataset through the proposed augmentations, the performance of the proposed VQG increases significantly, yielding a BLEU score of 39.74% and 11.6% on the validation set and the test set respectively.

An additional manual evaluation of the VQG models' outputs was performed by two experts in medical informatics. Twenty (question, image, category) triples from the test set were randomly selected for the manual evaluation. Detailed guidelines for the raters are listed in Subsection 4.2.2. Inter-rater reliability was calculated on each of the 3 measures. F1-score for each measure is presented in Table 5. Most of the reliability scores are close to .50, which is considered satisfactory reliability [55]. Table 4 presents the results of the manual evaluation.

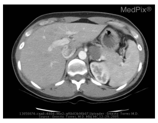
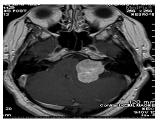
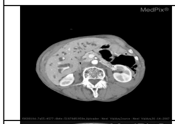
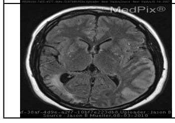
The human evaluation showed that our models achieved the highest scores by generating more relevant and correct questions. This also demonstrates that the image caption and the question category features contribute to generating better questions.

doi:10.20944/preprints202107.0385.v1

Table 3: Evaluation results on the validation set of the VQG dataset provided in the ImageCLEF 2020 VQA-Med challenge. VQGR trained on the original training datatset. VQGR$_{w\_our\_aug}$ trained on the augmented data obtained by our data augmentation technique.

Model	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**ROUGE-L**	**CIDEr**
VQGR	33.32	23.18	5.86	2.64	11.59	35.38	20.69
VQGR$_{w\_our\_aug}$	39.74	23.18	14.91	11.52	16.56	41.23	50.03

Table 6: Example image along with the question category, the automatically generated questions, and the ground truth question. The generated questions by our VQGRaD model and the baseline system are shown in blue and red, respectively. We manually selected the baseline's question from its outputs as the baseline system does not recognize the question category and generates a random question for each image.

Image	Category	Generated and ground truth questions
	Abnormality	is a ring enhancing lesion present in the right lobe of the liver? is a ring enhancing lesion present in the right lobe of the liver? is the liver normal ?
	Modality	was this mri taken with or without contrast ? which ventricle is compressed by the t2-hyperintense ? was this mri taken with or without contrast ?
	Organ	is this a typical liver ? are these normal laughed kidneys ? Is this a study of the brain?
	Plane	what plane is this image obtained ? what plane is this image blood-samples ? Is this image of a saggital plane?

Furthermore, the results showed that adding medical entities as an additional input improves the quality of the generated questions.

Overall, VQGRaD provides an improved approach to generating visual questions by targeting specific types of natural language questions about radiology images. Table 6 provides example questions generated by [19] (ground truth questions) and the VQGRaD model. These examples show that the questions generated by our model are more consistent with the reference questions.

The manual evaluation scores are much higher than the automatic ones. This is because the system, as shown in Table 6, generates the question words that are semantically comparable but does not generate the exact same words as the ground-truth answer. Indeed, we believe that the existing automatic evaluation metrics are not enough to accurately evaluate text/question generation tasks. Further efforts are needed to investigate a better evaluation strategy for the VQG task.

5. Conclusion and future work

In this paper, we presented a goal-driven visual question generation approach called VQGRaD that can generate a question that is relevant to the image and a specified category. In particular, we were interested in questions about Abnormality, Modality, Organ, and Plane of radiology images. The generated questions are evaluated using automatic and manual evaluations and are found to outperform the baseline systems. The manual evaluation showed that the generated questions appear comparable in quality to the human-generated questions. The results also showed that our data augmentation technique can boost performance on the VQG task.

Future work includes the creation of larger and more varied VQG datasets as well as the use of VQG models to create VQA data. We will also study additional question categories and investigate the use of the attention mechanism to focus on specific regions instead of the whole image. We also plan to investigate better evaluation strategies/metrics for the VQG task.

Acknowledgments: This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, X.; Grandvalet, Y.; Davoine, F.; Cheng, J.; Cui, Y.; Zhang, H.; Belongie, S.; Tsai, Y.H.; Yang, M.H. Transfer learning in computer vision tasks: Remember where you come from. *Image and Vision Computing* **2020**, *93*, 103853. doi:10.1016/j.imavis.2019.103853.
2. Guo, J.; He, H.; He, T.; Lausen, L.; Li, M.; Lin, H.; Shi, X.; Wang, C.; Xie, J.; Zha, S.; Zhang, A.; Zhang, H.; Zhang, Z.; Zhang, Z.; Zheng, S.; Zhu, Y. GluonCV and GluonNLP: Deep Learning in Computer Vision and Natural Language Processing, 2020, [arXiv:cs.LG/1907.04433].
3. Pelka, O.; Friedrich, C.M.; Garcıa Seco de Herrera, A.; Müller, H. Overview of the ImageCLEFmed 2020 concept prediction task: Medical image understanding. CLEF2020 Working Notes, CEUR Workshop Proceedings. CEUR-WS. org, Thessaloniki, 2020.
4. Sarrouti, M.; Alaoui, S.O.E. SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions. *Artificial Intelligence in Medicine* **2020**, *102*, 101767. doi:https://doi.org/10.1016/j.artmed.2019.101767.
5. Ruder, S.; Peters, M.E.; Swayamdipta, S.; Wolf, T. Transfer Learning in Natural Language Processing. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials; Association for Computational Linguistics: Minneapolis, Minnesota, 2019; pp. 15–18. doi:10.18653/v1/N19-5004.
6. Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; Hon, H.W. Unified Language Model Pre-training for Natural Language Understanding and Generation, 2019, [arXiv:cs.CL/1905.03197].
7. Moen, E.; Bannon, D.; Kudo, T.; Graf, W.; Covert, M.; Van Valen, D. Deep learning for cellular image analysis. *Nature methods* **2019**, pp. 1–14.
8. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019, [arXiv:cs.CL/1810.04805].
9. Ionescu, B.; Müller, H.; Villegas, M.; de Herrera, A.G.S.; Eickhoff, C.; Andrearczyk, V.; Cid, Y.D.; Liauchuk, V.; Kovalev, V.; Hasan, S.A.; others. Overview of ImageCLEF 2018: Challenges, datasets and evaluation. International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, 2018, pp. 309–334.
10. Pelka, O.; Friedrich, C.M.; Seco De Herrera, A.; Müller, H. Overview of the ImageCLEFmed 2019 concept detection task. CEUR Workshop Proceedings, 2019.
11. Ben Abacha, A.; Datla, V.V.; Hasan, S.A.; Demner-Fushman, D.; Müller, H. Overview of the VQA-Med Task at ImageCLEF 2020: Visual Question Answering and Generation in the Medical Domain. CLEF 2020 Working Notes; , 2020; CEUR Workshop Proceedings.
12. Gupta, D.; Suman, S.; Ekbal, A. Hierarchical deep multi-modal network for medical visual question answering. *Expert Systems with Applications* **2021**, *164*, 113993.
13. Mostafazadeh, N.; Misra, I.; Devlin, J.; Mitchell, M.; He, X.; Vanderwende, L. Generating Natural Questions About an Image. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Association for Computational Linguistics: Berlin, Germany, 2016; pp. 1802–1813. doi:10.18653/v1/P16-1170.
14. Zhang, S.; Qu, L.; You, S.; Yang, Z.; Zhang, J. Automatic Generation of Grounded Visual Questions. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), 2016, pp. 4235–4243, [arXiv:cs.CV/1612.06530].
15. Li, Y.; Duan, N.; Zhou, B.; Chu, X.; Ouyang, W.; Wang, X. Visual Question Generation as Dual Task of Visual Question Answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6116–6124. doi:10.1109/CVPR.2018.00640.
16. Krishna, R.; Bernstein, M.; Fei-Fei, L. Information Maximizing Visual Question Generation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2008–2018. doi:10.1109/CVPR.2019.00211.
17. Patro, B.N.; Kurmi, V.K.; Kumar, S.; Namboodiri, V.P. Deep Bayesian Network for Visual Question Generation. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1555–1565. doi:10.1109/WACV45572.2020.9093293.
18. Patil, C.; Patwardhan, M. Visual Question Generation: The State of the Art. *ACM Comput. Surv.* **2020**, *53*.
19. Lau, J.J.; Gayen, S.; Ben Abacha, A.; Demner-Fushman, D. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data* **2018**, *5*. doi:10.1038/sdata.2018.251.
20. Perez, L.; Wang, J. The Effectiveness of Data Augmentation in Image Classification using Deep Learning, 2017, [arXiv:cs.CV/1712.04621].
21. Inés, A.; Domínguez, C.; Heras, J.; Mata, E.; Pascual, V. Biomedical image classification made easier thanks to transfer and semi-supervised learning. *Computer Methods and Programs in Biomedicine* **2021**, *198*, 105782. doi:10.1016/j.cmpb.2020.105782.

22. Sarrouti, M.; Ben Abacha, A.; Demner-Fushman, D. Visual Question Generation from Radiology Images. Proceedings of the First Workshop on Advances in Language and Vision Research; Association for Computational Linguistics: Online, 2020; pp. 12–18.
23. Kalady, S.; Elikkottil, A.; Das, R. Natural language question generation using syntax and keywords. Proceedings of QG2010: The Third Workshop on Question Generation, 2010, Vol. 2, pp. 5–14.
24. Kim, Y.; Lee, H.; Shin, J.; Jung, K. Improving neural question generation using answer separation. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, Vol. 33, pp. 6602–6609.
25. Li, J.; Gao, Y.; Bing, L.; King, I.; Lyu, M.R. Improving Question Generation With to the Point Context. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, Vol. 33, p. 3216–3226, [1910.06036].
26. Serban, I.V.; García-Durán, A.; Gulcehre, C.; Ahn, S.; Chandar, S.; Courville, A.; Bengio, Y. Generating Factoid Questions With Recurrent Neural Networks: The 30M Factoid Question-Answer Corpus. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics: Berlin, Germany, 2016; p. 588–598, [arXiv:cs.CL/1603.06807].
27. Kafle, K.; Kanan, C. Visual Question Answering: Datasets, Algorithms, and Future Challenges. *Computer Vision and Image Understanding* **2017**, *163*, 3–20. doi:10.1016/j.cviu.2017.06.005.
28. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077–6086. doi:10.1109/CVPR.2018.00636.
29. Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C.L.; Batra, D.; Parikh, D. VQA: Visual Question Answering. 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2425–2433. doi:10.1109/ICCV.2015.279.
30. Goyal, Y.; Khot, T.; Agrawal, A.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *Int. J. Comput. Vision* **2019**, *127*, 398–414. doi:10.1007/s11263-018-1116-0.
31. Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Zitnick, C.L.; Girshick, R. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1988–1997. doi:10.1109/CVPR.2017.215.
32. Masuda-Mora, I.; Pascual-deLaPuente, S.; i Nieto, X.G. Towards Automatic Generation of Question Answer Pairs from Images. Visual Question Answering Challenge Workshop, CVPR 2016; , 2016.
33. Zhang, J.; Wu, Q.; Shen, C.; Zhang, J.; Lu, J.; van den Hengel, A. Goal-Oriented Visual Question Generation via Intermediate Rewards. In *Computer Vision – ECCV 2018*; 2018; pp. 189–204. doi:10.1007/978-3-030-01228-1_12.
34. Yang, J.; Lu, J.; Lee, S.; Batra, D.; Parikh, D. Visual Curiosity: Learning to Ask Questions to Learn Visual Recognition, 2018, [arXiv:cs.RO/1810.00912].
35. Jain, U.; Zhang, Z.; Schwing, A. Creativity: Generating diverse questions using variational autoencoders. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017; Institute of Electrical and Electronics Engineers Inc.: United States, 2017; pp. 5415–5424. doi:10.1109/CVPR.2017.575.
36. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* **2013**, [1312.6114].
37. Hasan, S.A.; Ling, Y.; Farri, O.; Liu, J.; Müller, H.; Lungren, M.P. Overview of ImageCLEF 2018 Medical Domain Visual Question Answering Task. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018; Cappellato, L.; Ferro, N.; Nie, J.; Soulier, L., Eds., 2018, Vol. 2125, *CEUR Workshop Proceedings*.
38. Ben Abacha, A.; Gayen, S.; Lau, J.J.; Rajaraman, S.; Demner-Fushman, D. NLM at ImageCLEF 2018 Visual Question Answering in the Medical Domain. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018; Cappellato, L.; Ferro, N.; Nie, J.; Soulier, L., Eds., 2018, Vol. 2125, *CEUR Workshop Proceedings*.
39. Ben Abacha, A.; Hasan, S.A.; Datla, V.V.; Liu, J.; Demner-Fushman, D.; Müller, H. VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019; Cappellato, L.; Ferro, N.; Losada, D.E.; Müller, H., Eds., 2019, Vol. 2380, *CEUR Workshop Proceedings*.
40. Al-Sadi, A.; Al-Theiabat, H.; Al-Ayyoub, M. The inception team at VQA-MED 2020: Pretrained VGG with data augmentation for medical VQA and VQG. CLEF, 2020.
41. Kobayashi, S. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. Proceedings of the 2018 Conference of the North American Chapter ofthe Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018. doi:10.18653/v1/n18-2072.
42. Şahin, G.G.; Steedman, M. Data Augmentation via Dependency Tree Morphing for Low-Resource Languages. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2018. doi:10.18653/v1/d18-1545.
43. Schmidhuber, J.; Hochreiter, S. Long short-term memory. *Neural Comput* **1997**, *9*, 1735–1780.
44. Bird, S.; Klein, E.; Loper, E. *Natural language processing with Python: analyzing text with the natural language toolkit*; " O'Reilly Media, Inc.", 2009.
45. Chen, X.; Fang, H.; Lin, T.Y.; Vedantam, R.; Gupta, S.; Dollár, P.; Zitnick, C.L. Microsoft COCO Captions: Data Collection and Evaluation Server. *ArXiv* **2015**, *abs/1504.00325*. doi:https://arxiv.org/abs/1504.00325.

46. Koehn, P.; Monz, C. Manual and automatic evaluation of machine translation between European languages. Proceedings of the Workshop on Statistical Machine Translation - StatMT '06. Association for Computational Linguistics, 2006. doi:10.3115/1654650.1654666.
47. Du, X.; Cardie, C. Harvesting Paragraph-level Question-Answer Pairs from Wikipedia. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 1907–1917. doi:10.18653/v1/P18-1177.
48. Hosking, T.; Riedel, S. Evaluating Rewards for Question Generation Models. Proceedings of the 2019 Conference of the North. Association for Computational Linguistics, 2019. doi:10.18653/v1/n19-1237.
49. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
50. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *ICLR* **2015**.
51. Aronson, A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the AMIA Symposium. American Medical Informatics Association, 2001, p. 17.
52. Sarrouti, M. NLM at VQA-Med 2020: Visual Question Answering and Generation in the Medical Domain. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020; Cappellato, L.; Eickhoff, C.; Ferro, N.; Névéol, A., Eds., 2020, Vol. 2696, *CEUR Workshop Proceedings*.
53. Al-Sadi, A.; Al-Theiabat, H.; Al-Ayyoub, M. The Inception Team at VQA-Med 2020: Pretrained VGG with Data Augmentation for Medical VQA and VQG. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020; Cappellato, L.; Eickhoff, C.; Ferro, N.; Névéol, A., Eds., 2020, Vol. 2696, *CEUR Workshop Proceedings*.
54. Hripcsak, G. Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association* **2005**, *12*, 296–298. doi:10.1197/jamia.m1733.
55. Viera, A.J.; Garrett, J.M.; others. Understanding interobserver agreement: the kappa statistic. *Fam med* **2005**, *37*, 360–363.