Original Article

Automatic Generation of Structured Radiology Reports for Volumetric Computed Tomography Images Using Question-Specific Deep Feature Extraction and Learning

Samira Loveymi, Mir Hossein Dezfoulian, Muharram Mansoorizadeh

Department of Computer Engineering, Bu-Ali Sina University, Hamedan, Iran

Abstract

Background: In today's modern medicine, the use of radiological imaging devices has spread at medical centers. Therefore, the need for accurate, reliable, and portable medical image analysis and understanding systems has been increasing constantly. Accompanying images with the required clinical information, in the form of structured reports, is very important, because images play a pivotal role in detect, planning, and diagnosis of different diseases. Report-writing can be exposure to error, tedious and labor-intensive for physicians and radiologists; to address these issues, there is a need for systems that generate medical image reports automatically and efficiently. Thus, automatic report generation systems are among the most desired applications. **Methods:** This research proposes an automatic structured-radiology report generation system that is based on deep learning methods. Extracting useful and descriptive image features to model the conceptual contents of the images is one of the main challenges in this regard. Considering the ability of deep neural networks (DNNs) in soliciting informative and effective features as well as lower resource requirements, tailored convolutional neural networks and MobileNets are employed as the main building blocks of the proposed system. To cope with challenges such as multi-slice medical images and diversity of questions asked in a radiology report, our system develops volume-level and question-specific deep features using DNNs. **Results:** We demonstrate the effectiveness of the proposed system on ImageCLEF2015 Liver computed tomography (CT) annotation task, for filling in a structured radiology report about liver CT. The results confirm the efficiency of the proposed approach, as compared to classic annotation methods. **Conclusion:** We have proposed a question-specific DNN-based system for filling in structured radiology reports about medical images.

Keywords: *Convolutional neural network, medical image analysis, MobileNet, radiology report generation*

Submitted: 31-Mar-2020 **Revised:** 20-Jun-2020 **Accepted:** 23-Sep-2020 **Published:** 21-Jul-2021

Introduction

Medical imaging plays an important role in the health care and treatment domains by facilitating rapid and reliable detection, treatment planning, and response analysis. With the advent of medical imaging systems and their widespread application, the number of medical images is continuously increasing. Most medical imaging techniques, which are employed in studying, diagnosing, and patient care planning, are noninvasive. This makes them less expensive than invasive methods and highly facilitates their usage by physicians and patients. Specialized medical professionals usually conduct the reading and interpretation of medical images via generating reports. For less experienced and even experienced radiologists, writing imaging reports is an inconvenient task as it is tedious, is time-consuming, and demands various skills.

This issue creates the need to build systems that automatically, reliably, and effectively process such data and generate useful and relevant information in the form of medical image reports. To make the generated reports more helpful for physicians and patients, they should be structured and domain specific. Such organized reporting with a standard format is useful in highlighting salient pieces of information, reflecting image interpretation, and elevating the efficiency

Address for correspondence:
Dr. Muharram Mansoorizadeh, Department of Computer Engineering, Bu-Ali Sina University, Hamedan, Iran.
E-mail: mansoorm@basu.ac.ir

Access this article online

Website: www.jmssjournal.net

DOI: 10.4103/jmss.JMSS_21_20

Quick Response Code:

How to cite this article: Loveymi S, Dezfoulian MH, Mansoorizadeh M. Automatic generation of structured radiology reports for volumetric computed tomography images using question-specific deep feature extraction and learning. J Med Sign Sens 2021;11:194-207.

of the workflow.[1] Furthermore, this results in increasing the search performance and retrieval rate from radiology databases for comparative diagnosis, medical education, etc., which are highly desirable in practice.

The automatic structured report generation system could be defined as a system that annotates medical images with clinical and specialized prior information, or a system that answers several clinical questions, given the patient's medical image. To this end, the medical image (e.g., radiology) automatic report generation involves conducting the following tasks: question answering,[2,3] image annotation,[4-6] or image classification tasks.[7-9] For the development of such systems, the main approaches in the literature are based on classic pattern recognition or modern deep learning methods.

Studies reveal that for implementing a computer-aided diagnosis system with purposes such as medical image analysis, and classification, the image representation, feature extraction, and learning methods play a pivotal role in the research efforts.[10]

Accordingly, the adopted image parameterization frameworks are grouped into two categories, namely classic handcrafted features extracted via engineered algorithms and modern learned features extracted via deep neural networks (DNNs).[11] The bulk of the literature in the structured radiology reports generation is based on medical image annotation or retrieval methods, which employ the classical approach. That is, extracting handcrafted features and training classification models on them. However, there is an expanding body of work, which adopts the modern approach: it takes the raw image signal as the input and leaves the feature extraction, to the DNN.

In this paper, we narrow down our scope to the liver ImageCLEF[12] annotation task in which the goal is to automatically fill in a structured radiology report about the liver computed tomography (CT) volumes. The main purpose of this task is to describe the physical and textural features of the liver, vessels, size of elements within the image, and the pathology of lesions. Classic approaches have been widely used along with this database. To that end, Kumar *et al*.[13] exploited content-based image retrieval methods to find similar images to the given query and annotated it using the tags of the retrieved images. They extracted classic engineered features such as shape, texture, pixel intensities, and scale-invariant feature transform-based bag of words and fed them into a two-stage support vector machine (SVM). They also utilized a weighted nearest neighbor classifier for annotation. In a study by Nedjar *et al*.,[5] for generating a structured radiology report for liver CT images, two methods were proposed which employ annotation techniques. In the first method, the shape and texture features are extracted using Haralick[14] and Gabor wavelets.[15] Then, random decision forests and nearest neighbor classifiers were trained on top of these features. In the second approach, they introduce the notion of liver image signature, which is synthesized from the liver images and uniquely represents each image, and using hamming distance function for retrieval. Second approach lead to achieving the best results in ImageCLEF2015 Liver CT Annotation task competition. In a study by Nedjar *et al*.,[6] bi-dimensional empirical mode decomposition texture features and Gabor wavelets are used for liver image annotation. The method incorporates intrinsic mode function decomposition to decompose an image into several components and extracts Gabor responses from the components. The labels of the input image are predicted using the labels of its five nearest neighbors. The study by Spanier *et al*.[16] focuses on the liver lesions and proposes an image retrieval system based on the annotations. In this method, SVMs, linear discriminant models, and logistic regression are used as binary classifiers and trained on the histogram of the gray levels (Hounsfield units) in the lesion regions. Loveymi *et al*.,[4] investigated the issue of the diversity of the question types and tried to alleviate it via learning question-specific models. Along with the classic shape and texture features, they introduced novel spatio-temporal texture features to address this issue.

In the above-cited work, the features are usually selected or engineered based on some priors and domain-specific knowledge. Then, a model structure is selected, and a proper training procedure is adopted to learn the model parameters/variables. An optimal feature should be as discriminant and robust as possible. To this end, the feature extraction process should filter out the irrelevant information and only pass through the task-related information. Underpinning such a filtering process in an abstract information space though a handcrafted pipeline of linear and nonlinear transforms is extremely challenging, if not impossible. In the modern deep learning-based approach, feature engineering is replaced with feature learning. The learning process allows for extracting task-specific features with the guidance of the objective function optimizer. This paves the way for passing through the useful information and discarding the task-irrelevant information. The learning instead of engineering-based feature extraction paradigm is at the heart of the state-of-the-art (STOA) pattern recognition techniques.[9,17] Hereafter, we refer to such features as deep features, which is an umbrella term for representations that are extracted from various types of DNN.

The success of the deep learning in various domains triggered an ever-increasing body of work and interest in the field of medical image analysis. In this regard, Jing *et al*.,[18] adopted a convolutional neural network-long short-term memory (CNN-LSTM)-based framework for automatic generation of medical imaging reports. They used a CNN for visual feature extraction and LSTM to model the long-term dependencies and generating long paragraphs in a report. Zhang *et al*.[19] organized a multimodal mapping

from medical images to symptomatic reports. They used an auxiliary attention sharpening module to detect the image–language alignments more efficiently. However, their generated symptomatic reports are limited to representing five types of cell appearance features, which makes their problem less complex than radiology report production. In a study by Xue *et al*.,[1] for generation of an automated radiology report, the CNNs are combined with the LSTM in a recurrent way. It is capable of generating high-level conclusive impressions and making specific descriptive findings sentence by sentence. Sugimori.[20] reports an effort on recognizing body parts from three-dimensional (3D) CT images. They use deep CNN networks to detect the brain, neck, chest, abdomen, and pelvis in the slices of the CT images. A specific dataset has been developed for each organ. AlexNet and GoogleNet[21] are another powerful CNN-based networks which are mainly used for image classification.[22,23] Baltruschat *et al*.[24] utilized ResNet as the multi-label classifier of the chest radiography images. Modality classification and concept detection in medical images is another application of deep learning which has been studied along with transfer learning. In a study by Singh *et al*.,[9] modality classification and concept detection in medical images using deep learning methods and transfer learning are studied. For modality classification, seven neural networks, namely VGG-16,[25] VGG-19,[25] ResNet-50,[26] Inception-v3,[27] Xception,[28] MobileNet,[29] and Inception-ResNet-v2,[30] have been investigated. All the networks are pretrained on ImageNet;[8] after pretraining, the lower layers are fixed and the last fully connected (FC) layer is trained from scratch to learn features tailored for the medical image analysis. Finally, a logistic regression classifier is trained on the outputs of the fully connected layer. Experimental results illustrated that the Inception-v3 and MobileNet have the highest recognition rates. In our earlier work,[31] we used fusion of deep and handcrafted features for generating structured radiology report. Deep features were extracted from the last layer of a fine-tuned MobileNet and handcrafted features using local binary pattern method.

Given the centrality of the DNNs in the STOA image-processing systems, in this study, we propose a deep learning-based framework for the automatic generation of structured radiology reports for volumetric medical images. The reasons for the selection of methods based on deep learning are as follows:

- Selecting the most informative low-level features for modeling high-level concepts is one of the main challenges in mining and analyzing the image data. Deep learning solves such a data representation problem through learning a cascade of linear and nonlinear transformations and with the aid of the objective function optimizer. The input is the raw image signal and the output is the corresponding label. Such end-to-end (raw data to a label) framework bypasses the need for any suboptimal feature engineering in the pipeline that may discard task-relevant information. As such, the end-to-end deep network would automatically learn and solicit the proper set of features from the raw image signal for the intended task. Furthermore, and in contrast to the classic methods where the front end is designed irrespective of the back end, the feature extracting and the classifier are merged in a unified structure and simultaneously optimized. Therefore, they best fit each other
- In general, the questions in radiology reports include several diverse topics from various specialties. This implies answering the questions, and writing report involves solving multiple problems; the optimal set of features for each one is different. While using an identical set of features for the different questions is suboptimal, the classic literature abounds with researches employ a shared set of selected handcrafted features for all kinds of questions.[5,6,12,13] This is unavoidable if one ought to apply classic features. To circumvent this problem, in our previous work,[4] we analyzed every question independently and found the most discriminative handcrafted features for each question (task) through feature selection methods. Although the first strategy (using the same set of features for all tasks) has lower training overhead as it does not involve any feature-filtering process, it leads to poorer overall performance. On the other hand, the DNNs allow for learning bespoke features for each task that are tailored for the given problem with a specific set of labels. This bypasses the need to select low-level features for every question manually which accompanies extra overhead and is prone to missing relevant information. Therefore, we take this route to learn the most discriminative and robust feature representation for each question.

Having decided to work with deep learning methods, there are three key issues to be addressed when designing and implementing such techniques:

- First, how should the input be presented to the network? Available options are two-dimensional (2D) color or intensity images, volumetric data, preprocessed set of regions of interest, and many more
- Second, each DNN architecture has its advantages and shortcomings; which architecture does best fit the application at hand?
- Finally, how much training data are available, and is it sufficient to build a reliable system? This is more critical when the architecture is large, which is the case in many deep learning applications. Possible solutions could be unsupervised pretraining,[32] data augmentation, and transfer learning. Hence, the question would be whether the training data are enough to train the model from scratch, or we should resort to the aforementioned solutions, for example, doing transfer learning via

pretrain the model on a larger database and only retrain the higher layers?

The optimal decision on these issues depends on the task, type of the images, and amount of the training data. In the current work, for generating a radiology report from volumetric abdominal CT images, we addressed the aforementioned challenges as follows:

- While the classification of the medical image using CNNs has achieved an eye-capturing success, it is still difficult to accurately characterize and classify volumetric medical images.[10] One of the main limitations is the fact that the optimization of the CNNs in the 3D (volumetric) classification tasks is not that straightforward. Dealing with this issue, we use deep networks with 2D inputs using slices of a volume as their input. Then, combine the extracted features of the slices of a volume image to produce the 3D or volume-level features
- The computational complexity of the DNNs could be very high depending on the utilized architecture. In medical applications, the techniques may be run on common desktop PCs, laptops, and even as apps on smartphones with limited computational capabilities and memory. Therefore, to maximize the applicability and usefulness of such technology in practice, they should demand minimal computational resources and memory footprint while preserving optimal performance. This puts an upper-bound on the architecture's size and consequently, its modeling capacity which, in turn, could hurt the performance. To address this challenge, we use the MobileNet and a CNN with tailored structure as the basic building blocks of our system. The conventional CNNs, owing to weight sharing, have fewer parameters compared with other architectures. The adopted MobileNet and CNN have even fewer parameters than the conventional CNNs, which makes them further suitable for our purpose[29,33]
- Finally, the DNN-based algorithms usually require a large amount of training data to render the expected high performance. However, in practice, and more specifically for many medical imaging tasks, for example, the subject of this work, large training datasets are not available. To address the challenge of training data size while benefiting from deep learning techniques, we take advantage of the data augmentation and transfer learning procedures in MobileNet. The features from the fine-tuned MobileNet and CNNs generate the slice-level features.

The rest of the paper is organized as follows. In the Methods Section, the proposed approach for structured radiology report generation is presented and explained in detail. The section Results is dedicated to the training procedure, hyper-parameter tuning, and deep feature extraction from MobileNet and CNN networks. The Discussion part includes the experimental results along with consultation. Finally, the Conclusion Section concludes the paper and puts forward some suggestions for future work.

Methods

In this paper, we aim at developing a learning system that automatically generates structured reports from the volumetric radiology images. The desired report is composed of multi-choice and binary questions about liver CT images and the pathology of lesions.

Dataset

We use Liver CT Annotation dataset from ImageCLEF, 2015,[12] that is specifically collected for structured radiology report generation on abdomen images. It contains questions on liver, its vessels, and pathology of lesions. The dataset includes information of fifty patients in the form of:

- Abdominal 3D CT image, as a cropped CT image of the liver 3D matrix. The volumes had various resolutions (x: 190–308 pixels, y: 213–387 pixels, slices: 41–588) and spacing (x, y: 0.674–1.007 mm, slice: 0.399–2.5 mm)
- Liver region, as a liver mask that specifies the part corresponding to the liver (a 3D matrix)
- Bounding box (volume of interest (VOI)) corresponding to the region of the selected lesion within the liver, as a vector of six numbers. They are corresponding to the coordinates of two opposite corners of the lesion's region within the CT volume
- Resource description framework file of the radiology report.

Figure 1 shows the architecture of the proposed system, which generates automatic structured radiology report for volumetric images using deep features. The questions in radiology reports of liver CT images are related to different heterogeneous parts such as the liver, lesion, and vessel. We put forward a claim that using the same feature set for answering all the questions is suboptimal. In other words, each question is a different problem with different decision boundaries, and the optimal set of features for solving it is different. This claim will be investigated and discussed in the Results section. Therefore, for each question/problem, one should extract the most informative and discriminative feature. Besides, a back end with sufficient modeling capacity should be employed to correctly and reliably generate report by answering the questions of the radiology questionnaire. As shown in Figure 1, the key novelty of the proposed system lies in the question-type guided feature extraction module, which is a DNN, trained based on the question type and the input training images.

Another challenge in the feature extraction stage is that the liver CT images are volumetric (3D), and it is important to pay attention to the multi-slice nature of volumetric images for making decisions and answering the questions. In this work, firstly, all 2D slices are treated as independent images. Slices contain detailed information on the lesions, hence

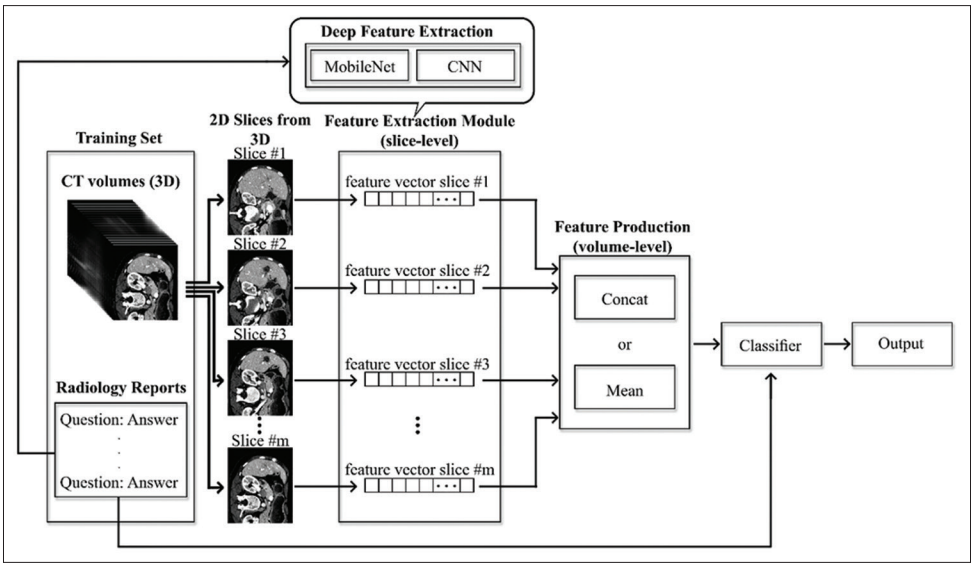

Figure 1: The architecture of the proposed structured radiology report generation model for volumetric images

slice-level learning is useful for extracting informative features. However, for answering questions in the report, prediction is conducted over the CT volumes. It is worth mentioning that in the slices selected for mining, liver and its lesions are all visible. Other abdominal elements were removed from the slices using the Liver Mask and Lesion VOI metadata in the dataset. Finally, it is important to follow the changing process of different parts of the CT image, both normal and abnormal, during the third dimension, and the CT volume will be considered as a sequence of 2D images (slices). Therefore, we will maintain the relationship between the content of different slices, via generating volumetric features by applying combination operators, such as concatenation and mean, on the slice-level features.

After feature extraction phase, binary or multi-class classifiers should be learned for the binary or multi-choice questions, respectively. In the proposed system, each question is tackled independently.

MobileNet

In this part of feature extraction module, we use MobileNet, a fully connected layer plus dropout function and a classification layer [Figure 2].[31]

MobileNet is a CNN with notably fewer number of parameters (size) and consequent complexity without considerable performance drop.[29] This architecture performs separable depth-wise convolution[29] and is instrumental in mobile and embedded vision systems. It has been shown that the ratio of the parameters of the MobileNet to the VGG-16, in the same level of accuracy, is 1:33.[34]

The main idea behind MobileNet is to replace the computationally intensive convolution layers with more computationally efficient depth-wise separable convolution. In this method, instead of applying convolution filter to all the input feature maps (channels), the convolution process is decomposed into two steps: first, a filter is applied to each channel (depth-wise convolution), and then, a 1 × 1 point-wise convolution combines all the feature maps.

The convolution layer is followed by a batch normalization[35] and rectified linear unit (ReLU).[36] The ReLU speeds up the training procedure, prevents gradient vanishing/explosions, and imposes sparsity. The activation function in MobileNet is ReLU6, as defined below, which is:

$$y = \min(\max(x, 0), 6) \quad (1)$$

Where x is the pixel value of the feature map and 6 is a hyperparameter which puts an upper bound (6, here) on the activation values.

As illustrated in Figure 2, the architecture of the MobileNet is composed of a set of blocks where each one consists of depth and point-wise convolutions, as well as batch normalization and activation functions (nonlinearity).

In this work, we used MobileNet-V2 in which each block is limited to three convolution layers. This structure maintains the low computational cost while rendering the required recognition accuracy.[33] Each block of the MobileNet-V2 starts with 1 × 1-convolution layer, which aims to increase effective image channels. This layer can be thought as the reverse of the projection layer. Expansion factor of the network controls the number of the resulting channels. Another innovation in the MobileNet-V2 is the residual (skip) connections in the blocks with the same number of input and output channels. Residual links improve the flow of the gradients in a deep network and alleviate the gradient vanishing issue. Version2 of the MobileNet has 17 blocks, followed by a 1 × 1 convolution and max or mean pooling layers.[33] The network is also characterized by a parameter α which is a depth scalar and defines the number of channels in each layer.

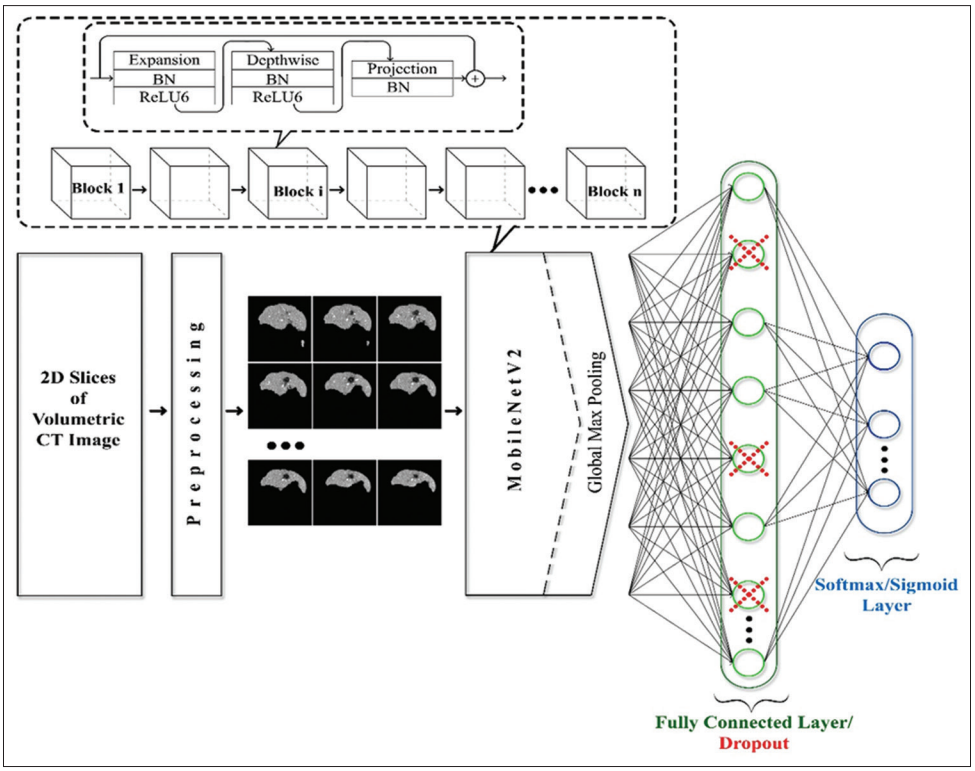

Figure 2: Learning a deep neural network for each question using MobileNet

Our goal is to build a network for each question and then extract features from the middle layers. In this regard, the MobileNet is followed by a FC layer to obtain a one-dimensional feature representation of the input image. The neurons of the FC layer are linked to all the neurons in the preceding global pooling layer. The largest share of the network parameters belongs to the FC layer. Using a narrow FC layer reduces the modeling capacity of the system and widening it elevates the liability to overfitting. To deal with this issue, we employ the dropout method: at each training step, the neurons are turned off (multiplied by 0) with probability 1 − p.[37] Dropout operation improves robustness via preventing complex co-adaptation between layers. The efficacy of dropout has been widely studied, for example, in a study by Warde-Farley *et al.*,[38] it is widely used in the STOA DNN-based systems as an efficient regularization technique.

Finally, the FC layer(s) activations are passed to a sigmoid or softmax output layer for binary or multi-class classification, respectively.

Convolutional neural networks

A CNN[39] is a special DNN architecture that contains a large number of fairly small convolution kernels (filters). The filter is characterized by the kernel weights which expectantly detect various task-desirable patterns in a given input image. Compared with fully connected layers, such structure keeps a number of trainable parameters remarkably low. In addition, using a CNN with a sufficiently large number of filters usually obviates the need for preprocessing steps such as affine normalization and/or region-of-interest bindings.[8]

A CNN is usually composed of convolution, pooling, and FC layers. The convolution layer contains pattern detectors, which operate over all the input locations. The max-pooling layer summarizes the output of the convolution layer and steers the focus to the most active units per input patch. Combined, these layers act as feature detectors that are nonlinear owing to ReLU activation function and *max* operation in pooling layer. Mathematically, we can write convolution layer as Eq. 2, and max-pooling layer as (3):

$$x_j^l = f\left(\sum_{i\in M_j} x_i^{l-1} * k_{ij}^l + b_j^l\right) \quad (2)$$

$$x_j^l = f\left(\beta_j^l max(x_i^{l-1} * k_{ij}^l) + b_j^l\right) \quad (3)$$

Where f (.) is a nonlinear function, x_j^l is the l^{th} layer j^{th} output, x_i^{l-1} is i^{th} input map from l to 1^{th} layer, *Mj* represents a selection of input maps, k and b are kernel and additive bias. β_j^l is a multiplicative bias of output feature map and finally, the size of the input map reduced through Eq. 3.

Furthermore, stacking several convolutions and pooling layers makes the feature detectors at the high levels, shift, and scale invariant, which is highly desirable in classification tasks.[40] The FC layer has similar functionality to the hidden layers in classic feed-forward networks. It

applies global transformations on the CNN's final feature maps, can adjust the dimension of the features (e.g., reduce dimension), extract more abstract representation, and make the features more linearly separable, paving the way for the linear classification takes place in the softmax layer.

Architecture of the proposed convolutional neural network

The adopted CNN in this work is similar to LeNet-5,[41] which is composed of two convolutional layers followed by three FC layers [Figure 3]. The first and second conv layers include six and 12 channels, respectively. For both layers, the stride, kernel, and pooling sizes were set to two, 5 × 5, and 2 × 2, respectively. ReLU activation has been applied at all layers except for the output layer where the logits turned into posterior probabilities via softmax function. Given the amount of training data, we avoided using larger networks which could run the risk of overfitting. Output of either of the FC layers may be used as an embedding or global feature vector for the given 2D image. We refer to such representations as deep features. Besides overfitting issue, to keep the size of the deep features in a comparable range with the classic features, we avoided using wider FC layers.

Deep features

As mentioned, after training the network, output of the internal layers can be deployed as feature representation for the given image (slice level). Here, we use the concatenation of the outputs of the last layer of the MobileNet and FC layers of CNN.

Results

Accuracy is the sole measure for evaluation and comparison of the developed approaches,[4-6,12,13,31] and the performance of the systems is evaluated via accuracy: the ratio of correctly labeled instances to all the instances. In our work, for each question that is about liver and its lesions or vessels, a specific model is learned. The system performance is measured as the mean accuracy over all the questions. To assess the generalization (out of sample) error, we have utilized cross-validation (CV). For comparison purposes, a classic machine learning model is also employed along with the DNN-based approaches. The model consists of SVM classifiers, which is learned on the top of the deep features of the volume images.

In this paper, only 43 of the questions have been considered for the task of ImageCLEF Liver CT Annotation2015. Questions with unbounded answers or no answer, for some images, are not included in the learning and evaluation processes.

Image preprocessing

Firstly, key-slices (slices where the lesion and liver are observed, they are selected using the Bounding Box VOI metadata in the dataset) are selected from every CT image. The number of slices in each CT image ranges from 30 to 500. In this study, we selected nine key-slices for each CT image in which lesion and liver are both visible. Finally, the liver is extracted through Liver Contour in each slice.

In deep feature extraction module, key-slices are fed into CNN and MobileNet-based DNN, which will be learned for every question independently. Figure 4 shows "Training_ Validation" loss functions of the CNNs, in the slice-level feature extraction phase for every individual question.

MobileNet configuration

The standard size of input images for MobileNet is 224 × 224 or 96 × 96, of which we selected the 96 × 96 for the image size. Transfer learning has been employed to pretrain the MobileNet. That is, the network has been pretrained on the ImageNet and fine-tuned on the dataset containing all the slices of volumes. Because the liver images are of gray scale, each image is repeated three times to imitate three channels of the RGB images. Binary and categorical cross-entropy is the loss functions for the binary and multi-label questions, respectively. To prevent overfitting, the initial learning rate is set to 10e-4 and decayed by factor of 0.9 after five epochs. The output of the last layer of the MobileNet layer has been extracted as the 1280-element

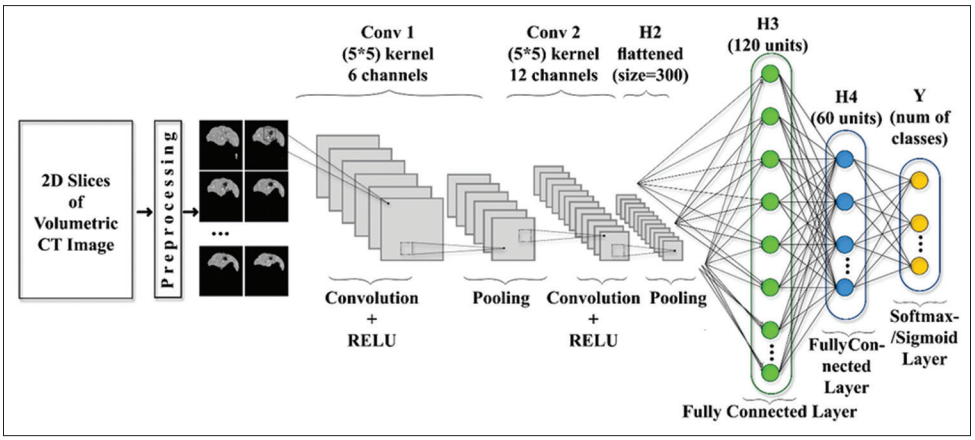

Figure 3: Proposed convolutional neural network architecture for the individual questions

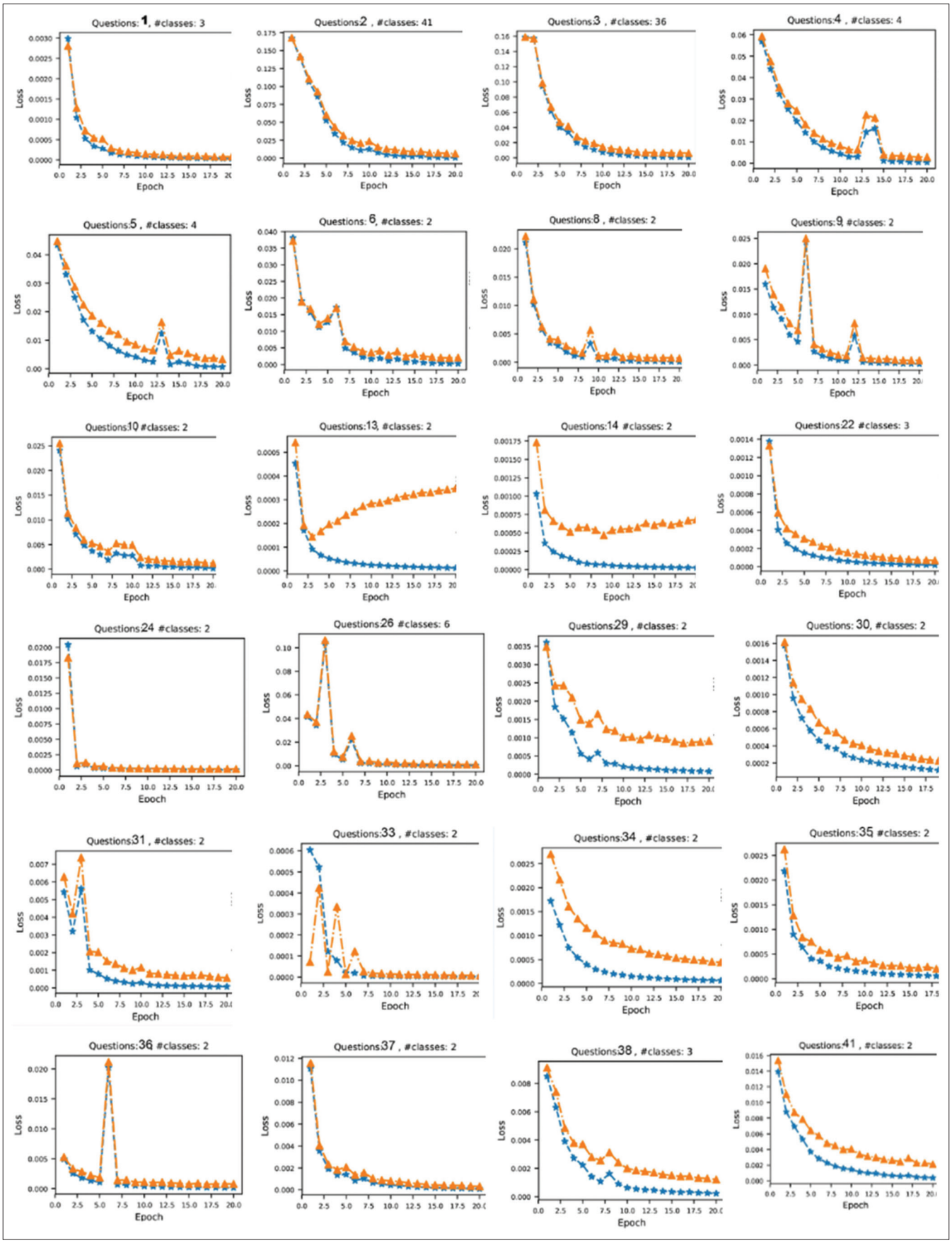

Figure 4: "Training_Validation" loss curves of the convolutional neural networks, in the slice-level feature extraction phase for every individual question

feature vector and reduced to 100 elements by the principal component analysis mapping. The network has been implemented and tested on Google colab1 machine with NVIDIA Tesla K80 GPU and 12 GB of memory.

Convolutional neural network configuration

Input images have been resized to 100 × 100 (we repeated our experiments with size 250 × 250 too, in which the accuracy was unchanged). The DNNs were trained from scratch using PyTorch.[42] All layers were initialized using the so-called Glorot method,[43] and cross-entropy has been employed as the loss function. We utilized SGD optimizer with a learning rate of 5e-4; to speed up the training, Nesterov momentum[44] with parameter 0.5 was applied and the weight decay (L2-norm) with parameter 1e-3 was used for regularization. We tried ADAM[45] and RMSProp[46] optimizers; both led to a slightly poorer performance for this task and architecture. From the aforementioned CNN system, the following four possible feature representations might be extracted: H3 preactivation (before ReLU), H3 activation (after ReLU), H4 preactivation, and H4 activation. The experimental results indicate the H4 activation features led to more accurate results than H3 features, although their size is half of the H3 layer. This is not surprising, as H4 activations represent a higher level of abstraction. The networks for this set of experiments have been implemented and tested on an NVIDIA GeForce GTX-1080Ti GPU.

Volume-level feature generation

In this study, for multi-slice analysis, deep features of key-slices are fused. To this end, all the slice-level deep features are averaged and concatenated for every volume. The accuracy of the proposed system for both mean and concatenation fusion operators is shown and compared in Table 1.

Classification

For most of the questions, the training data are imbalanced, that is, the number of samples per class is significantly different. Furthermore, some of the inspected classes are very close to each other, which could complicate the process of finding a reliable decision boundary. We applied SVM, which is a robust model due to max-margin criterion and also because it can work well in imbalance data scenarios. Moreover, it can cope well with the high dimensionality of the combined deep and handcrafted features. We employ the one-vs-all approach for multi-choice questions with linear kernels, which achieved the best accuracy among initial experiments using different kernels.

Table 1 shows the accuracy of the proposed systems along with the standard deviation for 10-fold CV. The accuracy is an average over all the 43 questions/problems. To be more precise, for every question, an SVM classifier is trained on question-specific learned deep features, the accuracy is computed per question [Table 2], and finally, the system's accuracy is the average across all the questions. Table 2 reports the detailed results; question number refers to the number in the rightmost column of Table 3.

Discussion

Table 1 summarizes the results of different question-specific volumetric deep features using CNN and MobileNet along with the SVM classifier with a linear kernel. As shown, the deep features extracted from tailored CNN and averaged across multi key-slices return better results than other deep features. Furthermore, CNN outperforms MobileNet with a margin of 10%, suggesting that more complex network would not essentially lead to higher performance. In our case, as an example of a small learning task, the relatively compact CNN architecture extracts features that are more informative and achieves higher accuracy for most of the questions.

Table 2 reports the recognition rates of three deep features for individual questions, compared with that of a study by Loveymi *et al*.,[4] which uses handcrafted features and finds the best feature set for every question, separately. Question# in Table 2 and Figure 5 refers to the number in the rightmost column of Table 3. Questions such as "Area-has-Area-Length-First," "Area-has-Area-Length-Second," and "Lesion-has-Lesion-Quantity" were excluded from the results reported for every question because they refer to the static size measurements and do not depend on the feature extraction method.

Another interesting observation is that the SVM with linear kernel outperforms other more complex nonlinear kernels, such as radial basis function. This is owing to the

1 https://colab.research.google.com

Table 1: Average of ten-fold cross-validation accuracy and standard deviation for all questions (using question-guided deep features)

Volume-level feature generator	Slice-level feature extractor	Dimension	Classifier	Accuracy (%)	Standard deviation
Mean of slices	CNN	60	SVM_Linear	97.76	1.41
	MobileNet	100	SVM_Linear	87.30	5.61
	CNN+MobileNet	60+100=160	SVM_Linear	97.71	1.39
Concatenation of slices	CNN	60×9=540	SVM_Linear	96.74	2.76
	MobileNet	100×9=900	SVM_Linear	84.47	5.04
	CNN + MobileNet	160×9=1440	SVM_Linear	96.56	2.44

CNN – Convolutional neural network; SVM – Support vector machine

fact that the output layer of a DNN, namely the softmax layer, is a linear classifier. Therefore, among others, the DNN layers should transform the features into a space in which the classes are linearly separable. This makes such features, that is, activation of higher layers, an optimal representation for SVMs with a linear kernel. In addition, note that optimizing SVMs with linear kernels is faster than their nonlinear counterparts, which makes this approach more appealing for practical scenarios.

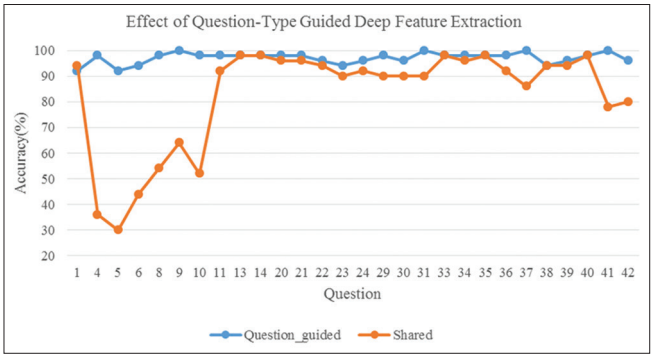

Figure 5: Accuracy obtained for every question with question-specific and shared feature set

In this work, a structured-radiology report generation system for volumetric medical images was proposed and tested on liver CT image task. The main contribution of this paper is question-guided deep feature extraction; this method has two advantages over previous work:[4]

1. Question-specific feature extraction: The hypothesis behind this idea is that questions in a radiology report are heterogeneous, about liver, lesion, and vessels, therefore low-level features to represent these parts may be different. As a matter of fact, each question is a different problem with distinguished decision boundaries. Therefore, the optimal set of features for fitting those boundaries would be different. If we use an identical shared feature set for all the questions like,[5,6,13] based on the propounded hypothesis, the corresponding system would be suboptimal. To prove this claim, we compared the performance of CNN that achieved the highest accuracy in Table 1, with the shared feature set scenario. The shared feature set means to optimize the DNN once, extract features from its internal layer, and use this feature set for all the questions. On the other hand, for the question-specific feature set, for every individual question, the DNN has optimized

Table 2: Accuracy obtained for every question with different question-guided deep features in comparison with a related work

Question number	CNN		MobileNet		CNN + MobileNet		Reference[4]
	Concatenation	Mean	Concatenation	Mean	Concatenation	Mean	
1	96	92	78	84	96	94	94
4	90	98	36	26	90	96	55
5	86	92	16	24	88	92	52
6	96	94	60	38	96	92	64
8	96	98	60	70	94	96	72
9	94	100	82	68	94	98	82
10	96	98	56	58	96	98	60
11	98	98	92	90	94	98	96
13	98	98	82	98	98	98	98
14	98	98	82	98	94	98	98
20	98	98	82	92	98	98	98
21	96	98	82	92	94	98	98
22	96	96	64	86	96	96	96
23	96	94	64	74	94	92	96
24	96	96	96	96	96	98	98
29	96	98	92	94	94	96	98
30	94	96	96	96	96	96	100
31	98	100	92	96	94	100	92
33	98	98	82	92	98	98	98
34	98	98	82	92	98	98	98
35	98	98	82	84	98	98	98
36	94	98	94	90	94	96	98
37	94	100	92	92	92	98	92
38	94	94	78	80	92	92	94
39	98	96	78	84	94	98	94
40	98	98	82	86	96	98	98
41	96	100	90	88	94	100	90
42	94	96	78	86	94	96	92

CNN – Convolutional neural network

Table 3: List of liver computed tomography structured radiology report questions and answers/annotations

Group	Concept	Properties	Values	Question number
Vessel	HepaticArtery	hasLumenDiameter	Decreased (0), increased (1), normal (2), other (3)	13
	HepaticArtery	hasLumenType	Obliterated (0), open (1), partially obliterated (2), other (3)	14
	HepaticPortalVein	hasLumenDiameter	Decreased (0), increased (1), normal (2), other (3)	15
	HepaticPortalVein	hasLumenType	Obliterated (0), open (1), partially obliterated (2), other (3)	16
	HepaticPortalVein	isCavernousTransformationObserved	NA (−1), true (1), false (0)	17
	HepaticVein	hasLumenDiameter	Decreased (0), increased (1), normal (2), other (3)	18
	HepaticVein	hasLumenType	Obliterated (0), open (1), partially obliterated (2), other (3)	19
	LeftHepaticVein	hasLumenDiameter	Decreased (0), increased (1), normal (2), other (3)	20
	LeftHepaticVein	hasLumenType	Obliterated (0), open (1), partially obliterated (2), other (3)	21
	LeftPortalVein	hasLumenDiameter	Decreased (0), increased (1), normal (2), other (3)	23
	LeftPortalVein	hasLumenType	Obliterated (0), open (1), partially obliterated (2), other (3)	24
	LeftPortalVein	isCavernousTransformationObserved	NA (−1), true (1), false (0)	25
	MiddleHepaticVein	hasLumenDiameter	Decreased (0), increased (1), normal (2), other (3)	34
	MiddleHepaticVein	hasLumenType	Obliterated (0), open (1), partially obliterated (2), other (3)	35
	RightHepaticVein	hasLumenDiameter	Decreased (0), increased (1), normal (2), other (3)	38
	RightHepaticVein	hasLumenType	Obliterated (0), open (1), partially obliterated (2), other (3)	39
	RightPortalVein	hasLumenDiameter	Decreased (0), increased (1), normal (2), other (3)	41
	RightPortalVein	hasLumenType	Obliterated (0), open (1), partially obliterated (2), other (3)	42
	RightPortalVein	isCavernousTransformationObserved	NA (−1), true (1), false (0)	43
Liver	LeftLobe	hasSizeChange	Decreased (0), increased (1), normal (2), other (3)	22
	RightLobe	hasSizeChange	Decreased (0), increased (1), normal (2), other (3)	40
	CaudateLobe	hasSizeChange	Decreased (0), increased (1), normal (2), other (3)	12
	Liver	hasDensity	Heterogeneous (0), homogeneous (1), other (2)	29
	Liver	hasLiverContour	Irregular (0), lobulated (1), nodular (2), regular (3), other (4)	30
	Liver	hasLiverDensityChange	Decreased (0), increased (1), normal (2), other (3)	31
	Liver	hasLiverPlacement	Downward displacement (0), normal placement (1), leftward displacement (2), upward displacement (3), other (4)	32
	Liver	hasSizeChange	Decreased (0), increased (1), normal (2), other (3)	33
Lesion	Lesion	hasLesionQuantity	1 (1), 2 (2), 3 (3), 4 (4), 5 (5), multiple (6)	26
	Lesion	LesionisDebrisObserved	True (1), false (0), NA (−1)	27
	Lesion	LesionisLevelingObserved	True (1), false (0), fluid (0), fluid gas (1), fluid solid (2), gas solid (3), other (4)	28
	Parenchyma	hasDensity	Heterogeneous (0), homogeneous (1), other (2)	36
	Parenchyma	hasParenchymaDensityChange	Decreased (0), increased (1), normal (2), other (3)	37
	Area	hasAreaDensity	NA (−1), hyperdense (0), hypodense (1), isodense (2), other (3)	1
	Area	hasAreaLengthFirst	A number in millimeter which represents the width of the lesion	2
	Area	hasAreaLengthSecond	A number in mm which represents the width of the lesion	3
	Area	hasAreaMarginType	Geographical (0), ill defined (1), irregular (2), lobular (3) serpiginous (4), speculative (5), well defined (6), other (7)	4
	Area	hasAreaShape	Band (0), fusiform (1), irregular (2), linear (3), nodular (4), ovoid (5), round (6), serpiginious (7), other (8)	5
	Area	hasDensityType	NA (−1), heterogeneous (0), homogeneous (1), other (2)	6
	Area	isCalcified	True (1), false (0), NA (−1)	7
	Area	isCentralLocalized	True (1), false (0)	8
	Area	isGallbladderAdjacent	True (1), false (0)	9
	Area	isPeriphericalLocalized	True (1), false (0)	10
	Area	isSubcapsularLocalized	True (1), false (0)	11

NA – Missing value

independently and then the features for every question are extracted from the corresponding DNN's internal layer. Experimental results in Table 4 show that the proposed question-specific deep feature set outperforms the shared deep features. As explained earlier, this is due to the fundamental differences between the

Table 4: Compare average accuracy using question-specific and shared feature set

Deep feature extraction (slice level)	Feature generator (volume level)	Dimension	Classifier	Accuracy (%)	Standard deviation
Question-type guided CNN	Mean of slices	60	SVM (linear)	97.76	1.41
Shared CNN	Mean of slices	60	SVM (linear)	87.53	2.06

CNN – Convolutional neural network; SVM – Support vector machine

Table 5: Accuracy of the proposed method and the results presented by others on the ImageCLEF Liver Computed Tomography Annotation 2015 dataset

References	Feature	Dimension	Classifier	Accuracy (%)
[13]	SIFT	1000	Weighted nearest neighbor	88.7
[5]	GLCM + 3D Gabor	111	Random forest	84.0
[6]	Gabor of BIMFs	960	CBIR/majority voting	88.9
[4]	DLBP/BoW/shape	36-500	SVM/random subspace KNN	93.1
Proposed method	CNN	60	SVM	97.76

CNN – Convolutional neural network; SVM – Support vector machine; SIFT – Scale-invariant feature transform; GLCM – Gray level co-occurrence matrix; CBIR – Content-based image retrieval; KNN – k-nearest-neighbor; BIMF – Bi-dimensional intrinsic mode functions; DLBP – Distance local binary pattern; BoW – Bag of visual words; 3D – Three dimensional

questions, which prevents a single set of features to be optimal for all questions/problems. Figure 5 compares the accuracies for different questions using shared and question-guided deep CNN-based feature sets

2. Applying deep features: Effectively addresses the data representation challenge and paves the way for finding the most informative and discriminative set of features for a given task. End-to-end networks that learn bespoke and efficient mappings for different questions result in higher accuracy than cited work which uses handcrafted features.[4] Table 5 demonstrates the higher accuracy of the proposed system. It is compared with the highest reported performance using handcrafted features for imageCLEF liver CT annotation dataset.

Results in Table 5 reveal that the proposed system outperforms similar works with a similar purpose, both those using question-specific handcrafted feature set and those that use shared feature sets with high-dimensional and diverse low-level features. This corroborates the claim that the proposed question-guided deep feature extraction method provides a more discriminative and informative representation for filling in a radiology report about CT images.

Conclusions

This paper proposed a system for automatic radiology report generation that aims to enhance the productivity and efficiency of the medical workfellows and using radiological image-based diagnosis. The proposed system takes advantage of both the self-exploratory nature of the modern deep learning techniques and the robustness and efficacy of the classic machine learning algorithms such as SVMs.

Due to the inherent problem of availability of small samples in medical domains, the proposed system utilized MobileNet and tailored CNN as the building blocks of its deep learning module. The networks follow the main idea of convolutional networks but have remarkably fewer parameters. Furthermore, the transfer learning has been deployed by pre-training the MobileNet on the ImageNet. Then, the network was fine-tuned using the small sample of the specific task.

In the proposed system, two end-to-end networks were implemented and trained for each question in a radiology report. That is, each question has its own set of deep features, extracted from its specialized deep network. Due to the variations in the sample distributions across classed for each question (imbalanced training data), and the sample size problem, SVM was employed as a robust classifier to mitigate these challenges. The proposed methods have been evaluated on the ImageCLEF2015 liver CT annotation dataset.[12] Question-specific deep features using CNN, lead to a more efficient learning system with higher accuracy. The proposed approaches can be readily applied in other imaging tasks and applications.

As future work, we plan to employ data augmentation procedures to synthetically expand the dataset, allowing to apply larger DNNs with higher modeling capacity. Another direction could be inspecting the trained network for possible knowledge distillations. Investigating the so-called tiny DNNs is also recommended for future work.

Financial support and sponsorship

None.

Conflicts of interest

There are no conflicts of interest.

References

1. Xue Y, Tao X, Rodney LL, Zhiyun X, Sameer A, Thoma Gr, *et al*. Multimodal recurrent model with attention for automated radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer; 2018.
2. Lau JJ, Gayen S, Ben Abacha A, Demner-Fushman D. A dataset

of clinically generated visual questions and answers about radiology images. Sci Data 2018;5:180251.
3. Wang, J, Hairong L, Rui J, Zhen X. Rule-Based Method to Develop Question-Answer Dataset from Chest X-Ray Reports. In: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). Cordoba, Spain: IEEE; 2019.
4. Loveymi S, Dezfoulian MH, Mansoorizadeh M. Generate structured radiology report from CT images using image annotation techniques: Preliminary results with liver CT. J Digit Imaging 2020;33:375-90.
5. Nedjar, I, Mahmoudi S, Chikh MA, Abi-Yad Kh, Bouafia Z. Automatic Annotation of Liver CT Image: ImageCLEFmed 2015. CLEF (Working Notes); 2015.
6. Nedjar I, Mahmoudi S, Chikh MA. Content-based Medical Image Tetrieval for Liver CT Annotation. Transactions on Machine Learning and Artificial Intelligence; 2017. p. 5.
7. Camlica Z, Tizhoosh HR, Khalvati F. Medical image classification via SVM using LBP features from saliency-based folded data. In: 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). Miami, Florida, USA: IEEE; 2015.
8. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks.In: Advances in Neural Information Processing Systems(NIPS) Nevada,USA; 2012.
9. Singh S, Ho-Shon K, Karimi S. Hamey L. modality classification and concept detection in medical images using deep transfer learning. In: 2018 International Conference on Image and Vision Computing New Zealand (IVCNZ). IEEE; 2018.
10. Afshar P, Mohammadi A, Plataniotis KN, Oikonomou A, Benali H. From handcrafted to deep-learning-based cancer radiomics: Challenges and opportunities. IEEE Signal Processing Magazine 2019;36:132-60.
11. Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE. A survey of deep neural network architectures and their applications. Neurocomputing 2017;234:11-26.
12. Marvasti NB, García MD, Üsküdarli S, Montes JF, Acar B. Overview of the ImageCLEF 2015 liver CT annotation task. In: CLEF (Working Notes). Workshop Proceedings. CEUR-¬WS. org, no. 1613-¬0073, Toulouse, France; 2015.
13. Kumar A, Dyer S, Kim J, Li C, Leong PH, Fulham M, *et al*. Adapting content-based image retrieval techniques for the semantic annotation of medical images. Comput Med Imaging Graph 2016;49:37-45.
14. Haralick RM, Shanmugam K, Dinstein IH. Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics; 1973. p. 610-21.
15. Lee TS. Image representation using 2D Gabor wavelets. IEEE Trans Pattern Analysis Mach Intellig 1996;18:959-71.
16. Spanier AB, Caplan N, Sosna J, Acar B, Joskowicz L. A fully automatic end-to-end method for content-based image retrieval of CT scans with similar liver lesion annotations. Int J Comput Assist Radiol Surg 2018;13:165-74.
17. Liu M, Zhang J, Nie D, Yap PT, Shen D. Anatomical landmark based deep feature representation for MR images in brain disease diagnosis. IEEE J Biomed Health Inform 2018;22:1476-85.
18. Jing B, Xie P, Xing E. On the automatic generation of medical imaging reports. arXiv:1711.08195v3:2017.
19. Zhang Z, Xie Y, Xing F, McGough M, Yang L. Mdnet: A Semantically and Visually Interpretable Medical Image Diagnosis Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017.
20. Sugimori H. Classification of Computed Tomography Images in Different Slice Positions Using Deep Learning. J Healthc Eng 2018;2018:1753480.
21. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, *et al*. Going Deeper with Convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015.
22. Holland L, Wei D, Olson KA, Mitra A, Graff JP, Jones AD, *et al*. Limited number of cases may yield generalizable models, a proof of concept in deep learning for colon histology. J Pathol Inform 2020;11:5.
23. Chen YC, Hong DJ, Wu CW, Mupparapu M. The use of deep convolutional neural networks in biomedical imaging: A review. J Orof Sci 2019;11:3.
24. Baltruschat IM, Nickisch H, Grass M, Knopp T, Saalbach A. Comparison of deep learning approaches for multi-label chest X-Ray classification. Sci Rep 2019;9:6381.
25. Simonyan K, Zisserman A, Very deep convolutional networks for large-scale image recognition. In Proc. International Conference on Learning Representations http://arxiv.org/abs/1409.1556 :2014.
26. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 770-8.
27. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016.
28. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017.
29. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, *et al*. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 2017.
30. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning. In: Thirty-First AAAI Conference on Artificial Intelligence; 2017.
31. Loveymi S, Dezfoulian MH, Mansoorizadeh M. Generate structured radiology report from liver CT images using fusion of mobilenet and local binary pattern. J Mach Vision Image Proc Forthcoming 2020;(Published Online, In Persian).
32. Erhan D, Courville A, Bengio Y, Vincent P. Why does unsupervised pre-training help deep learning? In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 210: p. 201-8.
33. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. Mobilenetv2: Inverted Residuals and Linear Bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018.
34. Li Y, Huang H, Xie Q, Yao L, Chen Q. Research on a surface defect detection algorithm based on MobileNet-SSD. Applied Sci 2018;8:1678.
35. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167v3 2015.
36. Dahl GE, Sainath TN, Hinton GE. Improving deep neural networks for LVCSR using rectified linear units and dropout. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: IEEE; 2013.
37. Baldi P, Sadowski PJ. Understanding Dropout. In: Advances in Neural Information Processing Systems; 2013.
38. Warde-Farley D, Goodfellow IJ, Courville A, Bengio Y. An empirical analysis of dropout in piecewise linear networks. arXiv preprint arXiv:1312.6197 2013.

39. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT press; 2016.
40. Garcia-Gasulla D, Parés F, Vilalta A, Moreno J, Ayguadé E, Labarta J. On the behavior of convolutional nets for feature extraction. J Artif Intellig Res 2018;61:563-92.
41. LeCun Y. LeNet-5, Convolutional Neural Networks. URL. Available from: http://yann.lecun.com/exdb/lenet/. [Last accessed on 2020 Oct 06].
42. Ketkar N, Santana E. Deep Learning with Python. Berkeley, CA: Apress; 2017.
43. Glorot X, Bengio Y. Understanding the Difficulty of Training Deep Feed for Ward Neural Networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics; 2010.
44. Sutskever I, Martens J, Dahl G, Hinton G. On the Importance of Initialization and Momentum in Deep Learning. In: International Conference on Machine Learning; 2013.
45. Kingma DP, Ba J. Adam: A method for stochastic optimization. In ICLR, 2015.
46. Tieleman T, Hinton G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA. Neural Network Mach Learn 2012;4:26-31.