IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

TECHO LAB: IMAGE CAPTIONING IN CHEST X-RAY BY USING CNN AND LSTM

[1]**Adithya E P**, [2]**Rinsha Shafrin T K**
Computer science and Engineering
METS SCHOOL OF ENGINEERING,MALA
Thrissur,India

***ABSTRACT*:** TECHO LAB: IMAGE CAPTIONING IN CHEST X-RAY BY USING CNN AND LSTM. Image captioning is a process of automatically generating a text description of an image, which is then used for reporting purposes. Image captions are a powerful tool for clinicians, allowing them to quickly identify a variety of diagnoses and conditions. And the use of image captioning in medical imaging has increased substantially over the past decade. The steadily increasing number of medical images places a tremendous burden on doctors who tend to read and write reports. If an image captioning model could generate drafts of the reports from the corresponding images, the workload of doctors would be reduced, thereby saving time and expenses. The contents of the X-ray image are predicted by words. The network of CNN and LSTM is used in this project. CNN is used to extract features, while LSTM is used to store the words one by one and make a sentence. The caption should not only be able to describe the disease but also make a sensible sentence that describes the Severity of that disease.

KEYWORDS: Chest x-ray, Image Captioning, CNN (Convolutional Neural Networks), LSTM (Long Short-Term Memory).

1 INTRODUCTION

Image captioning aims to describe the objects, actions, and details present in an image using natural language. Most image captioning research has focused on single-sentence captions, but the descriptive capacity of this form is limited; a single sentence can only describe in detail a small aspect of an image. Recent work has argued instead for image paragraph captioning to generate a (usually 5-8 sentence) paragraph describing an image. Compared with single-sentence captioning, paragraph captioning is a relatively new task. The main paragraph captioning dataset is the Visual Genome corpus, introduced by Krause et al.(2016). When strong single-sentence captioning models are trained on this dataset, they produce repetitive paragraphs that are unable to describe diverse aspects of images. The generated paragraphs repeat a slight variant of the same sentence multiple times, even when beam search is used.

Automatic radiology report generation is a computer-aided diagnostic technology used for generating a free-text description of disease diagnosis or future treatment based on radiology images (such as chest x-rays). Compared with general disease diagnosis technology, it is closer to artificial intelligence (AI), for it can not only output a list of numbers corresponding to the probabilities of possible diseases but also ‘‘write’’ an easy-to-understand report with natural language. With this technology, patients can read the chest X-rays by themselves, and no longer have to queue up to consult doctors. Moreover, the workload of radiologists will be greatly lightened.

Chest X-rays are the most common type of radiology image, which produces images of the heart, lungs, airways, blood vessels, and bones of the spine and chest, and is used for diagnosis and treatment of chest diseases, such as pneumonia and pneumothorax. A similar study area is a natural image captioning in computer vision and natural language processing because it has the same objective of mapping from images to text sequences. Hence, some common points exist between the two studies. First, encoder-decoder architecture is the basic architecture used to tackle these problems, in which the encoder, composed of a deep convolutional neural network (CNN), encodes images into a contextual vector, and the decoder, composed of long short-term memory (LSTM), decodes the contextual vector into a word sequence step by step.

1.2 Artificial Intelligence

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving. Artificial Intelligence is doing great progress in all factors, so automatically detecting the content of an image is a basic and important problem in AI fields that deals with computer sight and natural language processing. A deep recurrent architecture that interacts with the recent advances in done in image captioning in computer sectors and machine translation fields and that can be used to produce natural sentences that give detailed information about an image. Image captioning is a piece of work that requires the understanding of images and the awareness of producing correct description sentences with proper and suitable structure by extracting the features of an image. In this study, we try to understand a hybrid system describing the use of a Convolutional Neural Network (CNN) to generate an accurate description of the images and make use of an LSTM to accurately arrange the meaningful sentences using the removed or extracted keywords. CNN checks the similarity in the target image with a large dataset of training images and tries to generate an accurate description using the trained captions.

1.3 Machine Learning

Machine Learning is the idea to learn from examples and experiences, without being explicitly programmed. Instead of writing code, you feed data to the generic algorithm, and it builds logic based on the data given. Machine Learning is a field which is raised out of

Artificial Intelligence(AI). By applying AI, we wanted to build better and more intelligent machines. But except for a few mere tasks such as finding the shortest path between points A and B, we were unable to program more complex and constantly evolving challenges. There was a realisation that the only way to be able to achieve this task was to let the machine learn from itself. This sounds similar to a child learning from himself. So machine learning was developed as a new capability for computers. And now machine learning is present in so many segments of technology, that didn‘t even realise it while using it.

1.4 Deep Learning

Deep learning is a branch of machine learning which is completely based on artificial neural networks. Deep learning is an artificial intelligence function that imitates the workings of the human brain in processing data and creating patterns for use in decision-making. Deep learning is a subset of machine learning in artificial intelligence (AI) that has networks capable of learning unsupervised from data that is unstructured or unlabelled. It has a greater number of hidden layers and is known as deep neural learning or deep neural network. Deep learning has evolved hand-in-hand with the digital era, which has brought about an explosion of data in all forms and from every region of the world. This data, known simply as big data, is drawn from sources like social media, internet search engines, e-commerce platforms, and online cinemas, among others. This enormous amount of data is readily accessible and can be shared through fintech applications like cloud computing. However, the data, which normally is unstructured, is so vast that it could take decades for humans to comprehend it and extract relevant information. Companies realize the incredible potential that can result from unravelling this wealth of information and are increasingly adapting to AI systems for automated support. Deep learning learns from vast amounts of unstructured data that would normally take humans decades to understand and process. Deep learning utilizes a hierarchical level of artificial neural networks to carry out the process of machine learning. Artificial neural networks are built like the human brain, with neuron nodes connected like a web. While traditional programs build analysis with data in a linear way, the hierarchical function of deep learning systems enables machines to process data with a nonlinear approach.

1.5 Natural Language Processing

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken. NLP is a component of artificial intelligence (AI). The development of NLP applications is challenging because computers traditionally require humans to "speak" to them in a programming language that is precise, unambiguous and highly structured, or through a limited number of clearly enunciated voice commands. Human speech, however, is not always precise - it is often ambiguous and the linguistic structure can depend on many complex variables, including slang, regional dialects and social context.

1.6 Computer Vision

Computer vision tasks include methods for acquiring, processing, analyzing and understanding digital images, and extraction of high-dimensional data from the real world to produce numerical or symbolic information, e.g. in the forms of decisions. Understanding in this context means the transformation of visual images (the input of the retina) into descriptions of the world that make sense to thought processes and can elicit appropriate action. This image understanding can be seen as the disentangling of symbolic information from image data using models constructed with the aid of geometry, physics, statistics, and learning theory.

The scientific discipline of computer vision is concerned with the theory behind artificial systems that extract information from images. The image data can take many forms, such as video sequences, views from multiple cameras, multi-dimensional data from a 3D scanner, or medical scanning devices. The technological discipline of computer vision seeks to apply its theories and models to the construction of computer vision systems.

1.7 CNN

CNN stands for Convolutional Neural Networks. Convolutional networks are currently used in visual recognition. There are several convolutional layers in CNN. After these convolutional layers, text layers are fully connected layers as in a multilayer neural network. The CNN is designed in such a way that the benefit of the 2D structure of the input image can be taken. This target is accomplished with the help of several local connections and tied weights along with various pooling techniques which result in translation invariant features. The main advantages of using CNN are ease of training and possessing fewer parameters as compared to other networks with an equal number of hidden states.

CNN is used for extracting features from the image. CNN- Convolutional Neural networks are specialized deep neural networks which can process the data that has an input shape like a 2D matrix. Images are easily represented as a 2D matrix and CNN is very useful in working with images. CNN is used for image classifications and identifying if an image is a bird, a plane or Superman, etc. It scans images from left to right and top to bottom to pull out important features from the image and combines the feature to classify images. It can handle the images that have been translated, rotated, scaled and changed in perspective.

1.8 LSTM

LSTM stands for Long short-term memory; they are a type of RNN (recurrent neural network) which is well suited for sequence prediction problems. Based on the previous text, we can predict what the next word will be. It has proven itself effective from the traditional RNN by overcoming the limitations of RNN which had short-term memory. LSTM can carry out relevant information throughout the processing of inputs and with a forget gate, it discards non-relevant information. LSTM will use the information from CNN to help generate a description of the image.

2 Literature Review

[1] D. Hou, Z. Zhao, Y. Liu, F. Chang and S. Hu, "Automatic Report Generation for Chest X-Ray Images via Adversarial Reinforcement Learning," in IEEE Access, vol. 9, pp. 21236-21250, 2021, doi: 10.1109/ACCESS.2021.3056175.

1.Chest radiography, commonly called Chest X-ray (CXR), is a widely used radiology examination for the diagnosis of thoracic diseases. Radiological diagnosis requires abundant medical experience and a high level of concentration, which is challenging and time-consuming for radiologists.

In recent years, deep learning algorithms have made significant breakthroughs in image classification which are thus frequently utilized in medical diagnostics, However, multi-label classification on chest radiography remains a challenging task due to a variety of potential diseases contained in one scan. Some diseases are particularly hard to detect due to insufficient data and huge intra-class appearance variation, etc. Most of the previous work [5-6] treats the diseases independently because the common solution to multi-label classification is to treat it as multiple single-label problems, which is known as binary relevance. As a result, such models perform unsatisfactorily on hard-to-detect diseases, directly influencing the overall classification performance. This will automatically caption the X-ray.

2.An adversarial reinforced report-generation framework for chest X-ray images is proposed. Previous medical-report-generation models are mostly trained by minimizing the cross-entropy loss or further optimizing the common image-captioning metrics, such as CIDEr, ignoring diagnostic accuracy, which should be the first consideration in this area. Inspired by the generative adversarial network, an advert-serial reinforcement learning approach is proposed for the report generation of chest X-ray images considering both diagnostic accuracy and language fluency. Specifically, an accuracy discriminator (AD) and fluency discriminator (FD) are built that serves as the evaluators by which a report based on these two aspects is scored. The FD checks how likely a report originates from a human expert, while the AD determines how much a report covers the key chest observations. The weighted score is viewed as a ‘‘reward’’ used for training the report generator via reinforcement learning, which solves the problem that the gradient cannot be passed back to the generative model when the output is discrete. Simultaneously, these two discriminators are optimized by maximum-likelihood estimation for better assessment ability. Additionally, a multi-type medical concept fused encoder followed by a hierarchical decoder is adopted as the report generator. Experiments on two large radiograph datasets demonstrate that the proposed model outperforms all methods to which it is compared.

3.In this article, a novel medical report generation framework is proposed that considers both language fluency and diagnostic accuracy. From chest-radiograph images, the encoder extracts visual features and multi-type medical concepts, and then the hierarchical decoder inserts the medical concepts at the sentence and word levels to generate reports. More importantly, adversarial reinforcement learning (ARL) is introduced into the training procedure of medical report generation. The encoder-decoder is viewed as a generator and the reward modules as discriminators. In training iterations, discriminators are optimized by maximum-likelihood estimation, whereas the generator is trained by reinforcement learning.

Finally, the reward modules give highly accurate rewards and the generator generates better reports. In experiments, first, the high performance of the proposed full model is proved by performance comparison with several classical or recently proposed models from different aspects on two large chest X-ray datasets. Ablation studies are then conducted to verify the effectiveness of the language fluency discriminator (FD) and the diagnostic accuracy discriminator (AD), followed by trade-off parameter analysis and qualitative analysis. All of the experimental results demonstrate that the proposed fully learnable ARL architecture that combines AD and FD is superior to purely traditional optimization by cross-entropy alone, or to additional RL with manually designed reward functions.

[2] C. Amritkar and V. Jabade, "Image Caption Generation Using Deep Learning Technique," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697360.

1.A good description of an image is often said for ’Visualizing a picture in the mind’. The creation of an image in mind can play a significant role in sentence generation. Also, humans can describe the image after having a glance at it. The progress in achieving complex goals of human recognition will be done after studying existing natural image descriptions.

This task of automatically generating captions and describing the image is significantly harder than image classification and object recognition. The description of an image must involve not only the objects in the image but also the relation between the objects with their attributes and activities shown in images. This paper proposes a model capable of generating novel descriptions from images.

2.In Artificial Intelligence (AI), the contents of an image are generated automatically which involves computer vision and NLP (Natural Language Processing). The neural model which is regenerative is created. It depends on computer vision and machine translation. This model is used to generate natural sentences which eventually describe the image. This model consists of Convolutional Neural Network(CNN) as well as Recurrent Neural Network(RNN). CNN is used for feature extraction from images and RNN is used for sentence generation. The model is trained in such a way that if an input image is given to the model it generates captions which nearly describe the image. The accuracy of the model and smoothness or command of the language model learned from image descriptions are tested on different datasets. These experiments show that model is frequently giving accurate descriptions for an input image.

3.This work presents a model, which is a neural network that can automatically view an image and generate appropriate captions in a natural language like English. The model is trained to produce the sentence or description from a given image. The descriptions or captions obtained from the model are categorized into:

- Description without errors
- Description with minor errors
- Description somewhat related to the image
- Description unrelated to the image

The categories in results are due to the neighbourhood of some particular words, i.e., for words like the car it’s neighbourhood words like vehicle, van, cab etc. are also generated which might be incorrect. After so much of experiments, it is conclusive that the use of larger datasets increases the performance of the model. The larger dataset will increase accuracy as well as reduce losses. Also, it will be interesting how unsupervised data for both images, as well as text, can be used for improving the image caption generation approaches.

[3] K. C. Nithya and V. V. Kumar, "A Review on Automatic Image Captioning Techniques," 2020 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2020, pp. 0432-0437, doi: 10.1109/ICCSP48568.2020.9182105.

1.Ongoing advancement in Automatic Image Captioning (AIC) has demonstrated that it is conceivable to depict the most remarkable data passed on by pictures with exact and correct sentences. Image captioning means the creation of a textual depiction

of an image automatically. Captioning of images goes for portraying a picture utilizing characteristic language. The important step in image captioning is identifying different objects in an image, finding their relationships and classifying them and combining the words that may not use proper language modelling. Captioning with proper sentences requires computer vision and Natural Language Processing (NLP) for obtaining accurate sentences. Classified objects are then passed to the language model to create captions. Semantic knowledge about an object in a picture needs to obtain by capturing the characteristics of an image globally and locally. There are various methods used for captioning an image but supervised learning provides a better experience. This paper deals with some of the methods for image captioning.

2.Automatic Image Captioning (AIC) is one of the ongoing development fields. The image captioning technique involves object identification and modelling it into proper sentences by using NLPs. AIC helps to trace major salient characteristics in an image with error-free and proper sentences. Understanding a picture to a great extent relies upon acquiring picture highlights. The systems utilized for this reason can be extensively separated into traditional learning-based and deep learning-based strategies for AIC.
Initially, captioning has just endeavoured to yield straightforward portrayals for pictures taken under very compelled conditions. As a difficult and important research field, AIC is drawing in increasingly more consideration and is getting progressively significant. The captioning system needs a correlation between each element in an image along with actions and aspects. Generally, AIC tries to provide simple captions for a picture in any situation. Hence it is a difficult task to extract all the characteristics of an image along with proper NLPs. AIC have many applications. They can be used for blinds to identify images, by converting text into audio. They can be also used for indexing an image which is important for CBIR (" Content-Based Image Retrieval"). Image captioning also have various other purposes such as web searching, Facebook, education purpose etc.
But the challenge in image captioning is the large number of datasets required to get a meaningful dataset. Automatic image caption includes two steps. The first step is to identify individual objects in a picture by utilizing lines and strokes present in an image. Then separate the features into small meaningful parts. Various visual locales from which visual highlights are extricated. Then contrast it with the current database, which discovers the level of intelligibility between the components of the picture present and the database. The second part of the test includes utilizing key terms to shape sentences that caption the picture precisely. To make proper sentences, the translation of pictures to text is performed by using proper language.

3.Automatic image captioning is an emerging field in recent years. After reviewing some papers we found that various approaches like N-cut and colour-based segmentation, hybrid engine, and encoder-decoder framework are used. It mainly utilises computer vision along with natural language processing. Automatic image captioning using neural networks is a more advanced and accurate method. AIC age for pictures for individuals who experience the ill effects of different degrees of visual debilitation; the programmed making of metadata for pictures (ordering) for use via web indexes; universally useful robot vision frameworks; and numerous others. An enormous number of datasets are required. There are many open sources, for example, MSCOCO, and FLICKER datasets available. Changing the model engineering, for example, incorporating a consideration module and accomplishing more hyperparameters such as batch size number of layers and units etc improves the picture captioning framework.

[4] Y. Hu, Y. Zhang, T. Zhang, S. Gao and W. Fan, "Label Generation Network based on Self-selected Historical Information for Multiple Disease Classification on Chest Radiography," 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea (South), 2020, pp. 1015-1019, doi: 10.1109/BIBM49941.2020.9313507.

1.Deep learning has made significant breakthroughs in image classification, but accurate diagnosis on chest radiography remains challenging due to a variety of potential diseases contained in one scan. Complex relations among diseases have significant clinical meanings but are always ignored in most previous work. Thus in this paper, we propose a novel Label Generation Network (LGN) which treats the label sequence as the caption of a radiology image and utilizes RNN to generate the disease labels according to the semantic relations and co-occurrence dependency among them. However, the sequential generation process of RNN makes it hard to capture the complex topological relations among diseases. To mitigate this problem, a Historical Information Module (HIM) is specially introduced to LGN, in which all the generated labels are fully considered when generating a new label. Moreover, a specific self-attention mechanism is applied in HIM to learn the topological disease relations and utilize them to select useful historical information which can provide positive guidance to the prediction of the new labels. Very positive results have been obtained in our experiments on the benchmark dataset of Chest X-ray14, which significantly outperforms the state-of-the-art methods.

2.Chest radiography, commonly called Chest X-ray (CXR), is a widely used radiology examination for the diagnosis of thoracic diseases. Radiological diagnosis requires abundant medical experience and a high level of concentration, which is challenging and time-consuming for radiologists. In recent years, deep learning algorithms have made significant breakthroughs in image classification, which are thus frequently utilized in medical diagnostics, However, multi-label classification on chest radiography remains a challenging task due to a variety of potential diseases contained in one scan. Some diseases are particularly hard to detect due to insufficient data and huge intra-class appearance variation, etc. Most of the previous work treats the diseases independently because the common solution to multi-label classification is to treat it as multiple single-label problems, which is known as binary relevance. As a result, such models perform unsatisfactorily on hard-to-detect diseases, directly influencing the overall classification performance.

3.In this paper, we propose a novel Label Generation Network (LGN) for the multi-label classification of thoracic diseases on chest radiography. Different from previous work, LGN treats the disease label sequence as the caption of a CXR image and generates it with RNNs. Besides, the Historical Information Module (HIM) is introduced to LGN to capture the global relations among diseases and select useful historical information for the prediction of new labels. In our future work, one potential concern with LGN is the risk of learning biased relations and dependencies from a limited training set. The best solution is to explore some consistent known relational structures to model the relations with the guidance of professional medical knowledge. But the prediction of the entire caption, given the image does not happen at once. We will predict the caption word by word. Thus, we need to encode each word into a fixed-sized vector.

3 Project Discribtion

3.1Proposed System

The goal of image paragraph captioning is to generate descriptions from an image. This uses a hierarchical approach for text generation. Firstly, the objects in the image are detected and a caption related to that object is generated. Then combine the captions to get the output. Tokenization is the first module in this work where character streams are divided into tokens which are used in data(paragraph) preprocessing. It is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. The tokens are stored in a file and used when needed.

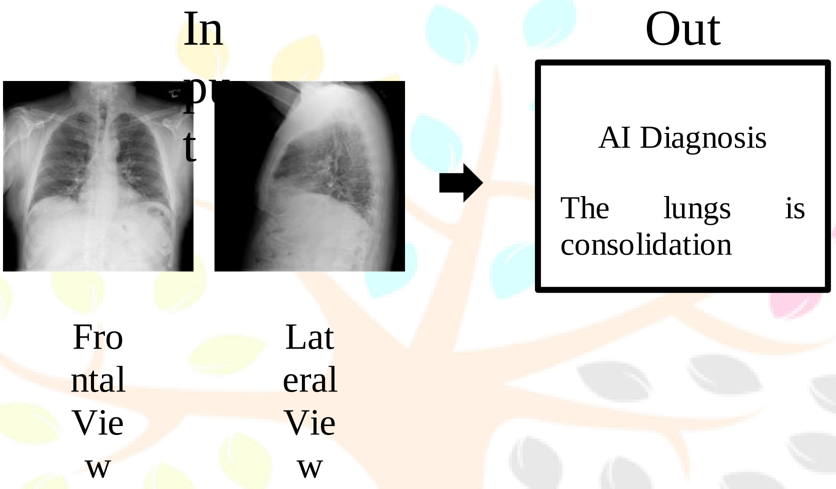

fig 3.1: input-output model

3.2 Purpose

The project goal is to come up with a caption for an X-ray image. Here image captioning is the process of making a description for an X-ray image. It necessitates an understanding of the important things, their characteristics, and the relationships between them. And image objects Deep learning techniques have progressed, and as a result, we can create models that can predict the future thanks to the availability of large datasets and computing power. In this project, our intended audiences are doctors and lab technicians. Doctors can use this website as an assistant to verify a patient's chest X-ray by this they can save their time. In the case of laboratories, they can provide the chest X-rays description along with the X-ray. And this will help to increase their profit. Here we are helping doctors and laboratories.

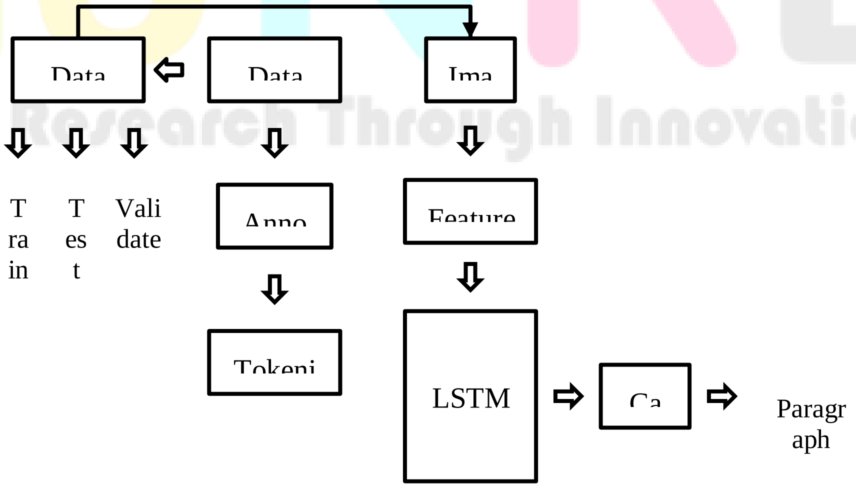

fig 3.2 system architecture

3.3 Scope of Project

There is no existing application for this. This application is reducing the load of doctors and lab technicians. There is no loss of time if they upload the image it will automatically generate the caption for the x-ray.

Advantages

- Less time-consuming.
- Easy, efficient and simple.
- Give quick and efficient services to the user.
- Fast and convenient.
- High-speed response to the user.
- If we insert an image within seconds it will provide output.

3.4 Overview of Document

The project techolab is a web application that has two users doctors and lab technicians. And the web application works like this, when uploading an x-ray there automatically generates the captions of the particular chest x-ray. The caption will be a meaningful sentence. Caption generation is a challenging AI problem where a textual description must be generated for a given image. here we are providing an X-ray as the input and a caption as the output. It will automatically generate the output for the given X-ray. Most image captioning systems use an encoder-decoder framework, where an input image is encoded into an intermediate representation of the information in the image, and then decoded into a descriptive text sequence.

CNN is as encoder and LSTM is the decoder. It is feasible to calculate automatic metrics efficiently. You won't have to waste time searching for captions because they'll be generated automatically.

4 CONCLUSION

This paper mainly focuses on X-ray captioning based on research papers. Different Captioning metrics are used for the evaluation of the sentences generated by the system. The scores tell about the accuracy of the words obtained. Different methods are compared which tells the efficiency of the LSTM method to be 80%. This provides the best results on Visual Genome Dataset. Hence, this paper provides a clear view of how a paragraph is generated from an image. The scope of the paper is limited to the LSTM approach only. In future, the scope of the work can be extended so that the system can be more efficiently used by all the researchers.

In future, the scope of the work can be extended so that the system can be more efficiently used by all the researchers.

REFERENCES

1) S. Hochreiter and J. Schmidhuber, ``Long short-term memory,'' Neural Comput., vol. 9, no. 8, pp. 17351780, 1997.
2) B. Jing, P. Xie, and E. Xing, ``On the automatic generation of medical imaging reports,'' 2017, arXiv:1711.08195. [Online]. Available: http://arxiv.org/abs/1711.08195
3) G. Liu, T.-M. Harry Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, and M. Ghassemi, ``Clinically accurate chest X-ray report generation,'' 2019, arXiv:1904.02633. [Online]. Available: http://arxiv.org/abs/1904.02633
4) Y. Xue and X. Huang, ``Improved disease classification in chest X-rays with transferred features from report generation,'' in Proc. Int. Conf. Inf. Process. Med. Imag. Cham, Switzerland: Springer, 2019, pp. 125138.
5) J. Yuan, H. Liao, R. Luo, and J. Luo, ``Automatic radiology report generation based on multi-view image fusion and medical concept enrichment,''in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Cham, Switzerland: Springer, 2019, pp. 721729.
6) A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ``Attention is all you need,'' in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 59986008.
7) K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, ``Show, attend and tell: Neural image caption generation with visual attention,'' in Proc. Int. Conf. Mach. Learn., 2015, pp. 20482057.