

## A Unified Visual Saliency Model for Automatic Image Description Generation for General and Medical Images

Sreela Sreekumaran Pillai Remadevi Amma<sup>\*1</sup>, Sumam Mary Idicula<sup>2</sup>

<sup>1</sup>Department of Computer Science, Government College Kariavattom, Thiruvananthapuram, Kerala, India

<sup>2</sup>Department of Computer Science, Muthoot Institute of Technology and Science, Kochi, Kerala, India

### ARTICLE INFO

Article history:

Received: 21 January, 2022

Accepted: 09 March, 2022

Online: 28 March, 2022

Keywords:

Image Description Generation

Image captioning

Deep learning

Visual Attention

### ABSTRACT

An enduring vision of Artificial Intelligence is to build robots that can recognize and learn the visual world and who can speak about it in natural language. Automatic image description generation is a demanding problem in Computer Vision and Natural Language Processing. The applications of image description generation systems are in biomedicine, military, commerce, digital libraries, education, and web searching. A description is needed to understand the semantics of the image. The main motive of the work is to generate description of the image using visually salient features. The encoder-decoder architecture with a visual attention mechanism for image description generation is implemented. The system uses a Densely connected convolutional neural network as an encoder and Bidirectional LSTM as a decoder. The visual attention mechanism is also incorporated in this work. The optimization of the caption is also done using a Cooperative game-theoretic search. Finally, an integrated framework for an automatic image description generation system is implemented. The performance of the system is measured using accuracy, loss, BLEU score and ROUGE. The grammatical correctness of the description is checked using a new evaluation measure called GCorrect. The system gives a the-state-of-art performance on the Flickr8k and ImageCLEF2019 challenge dataset.

## 1. Introduction

Computer vision researchers nowadays mainly focus on descriptive language to describe the world. Image description generation is a challenging topic in computer vision and natural language processing. Deep learning technology has produced tremendous progress in the automatic generation of image descriptions. The image description expresses the semantics and linguistic representation of the image. Image captioning applications include image retrieval, automatic video surveillance, image indexing, education, aid to visually impaired people, etc. The caption of a photo contains objects, attributes, spatial relationships, and actions. The description of an image should be meaningful, self-contained, grammatically, and semantically correct.

Generating image descriptions is an essential process in the area of both Computer Vision and Natural language processing. Imitating the human attitude for giving descriptions for images by a machine is a noticeable step along with the rapid growth of Artificial Intelligence. This task's significant challenge is to capture the relationships of objects in an image and generate a

natural language description. Traditionally, predefined templates are used for generating descriptions. However, this approach does not give enough variety available for creating lexically and semantically detailed descriptions. This limitation has been conquered with the increased efficiency of neural network models. Neural networks generate captions in state-of-the-art models by giving an image as input and forecasting the output description. The automatic image description generation system has many critical applications, such as aiding visually impaired people, building an intelligent robot, and making Google Image Search better than Google Keyword Search.

Connecting image and language is a complex problem in Computer Vision and Natural language Processing. Based on the literature, a comprehensive scene understanding is difficult. The image description systems should produce grammatically correct, relevant, human-like, and describe accurate information. For generating a better caption, the vital image features should be selected. Content selection from images is a significant problem. To optimize the description, the max search and beam search are commonly used methods for determining words. So, caption optimization needs to be improved to eliminate the limitations of

\* Corresponding Author Sreela S R, Email: [sreela148@gmail.com](mailto:sreela148@gmail.com)

max search and beam search.

Analyzing and summing up ideas from clinical pictures such as radiology images is a tedious task that specialists can handle. Automatic methods approximate the mapping from visual information to condensed textual descriptions. All medical images training data are accompanied by UMLS concepts extracted from the original image caption. Medical image captioning is an actual application of automatic image description generation. In this work, the proposed automated image description system is used for medical image captioning.

- We developed two components as a visual attention architecture and integrated automatic image description generation system to achieve the objectives. They are explained below:
- Visual attention: A hybrid architecture for visual attention is implemented. Spatial, channel-wise, and layer-wise attention are the components of visual attention.

Integrated automatic image description generation system: The image features are extracted using Densenet. The description is generated using Bidirectional Long Short term memory (BLSTM). The caption optimization is implemented using game-theoretic search. The framework has experimented on the Flickr8k dataset and medical image dataset ImageCLEF2019.

## 2. Related Works

Recent trends in Computer Vision and Natural language processing have influenced image description generation. An automatic image description generation is essential for many reasons, such as image understanding, image indexing, image searching, etc. Many research works have been progressed in image description generation in the last ten years. Image captioning systems are classified into Traditional machine learning-based systems and deep learning-based systems. In our literature, we concentrate more on deep learning-based approaches. The taxonomy of deep learning-based image description generation is based on six criteria: type of machine learning approach, the model architecture, feature mapping, the language models used, the number of captions produced, and others. The model architectures used in this system are encoder-decoder architecture and compositional architecture. The features are mapped into two spaces, such as visual space and multimodal space. The language models used in this system are Recurrent Neural Network(RNN) [1], Long Short Term Memory(LSTM) [2], etc. The captioning systems are divided into dense captioning and scene-based captioning based on the number of captions generated. Other image description generation systems are attention-guided, semantic concept-oriented, novel object-based, and stylized captions.

From the literature, the image description generation systems are classified as follows.

Direct Generation model: This model extracts the image's visual content and generates a description. Google's Neural Image caption generator [3], BabyTalk [4], Midge [5], Karpathy's system [6] follow this model.

Visual space model: The visual model finds the identical images of the query image and maps the description to the image.

Multimodal space model: This model finds similar images from multimodal space such as visual and textual. This kind of image description generation system is considered a retrieval problem.

The image captioning systems are further classified into template-based systems and Deep Neural Network-based systems.

Template-based: In this approach, the captions are generated using the objects, attributes, and scenes. Farhadi et al. The Markov random field, GIST, and Support Vector Machine(SVM) are used for caption generation and transform the scene contents into natural language sentences using the template in [7] systems. The Conditional Random Field (CRF) relates the objects, attributes, and prepositions systems [4]. Midge [5] systems generate text using the Berkeley parser and Wordnet ontologies. The disadvantage is that they produce inaccurate descriptions due to wrong object detection. Classical machine learning algorithms are used for object detection, which results in bad performance.

Deep Neural Network-based approach: The image captioning system involves image to text translation. Currently, the image captioning system consists of two parts: Encoder and Decoder. The encoder is used for extracting features of the picture. Deep neural networks are used for encoding, which has the highest accuracy in object categorization. The decoding module is realized using recurrent neural networks or LSTM, which is practiced for caption generation. The main components in [8] system are fully connected neural networks and multimodal log-bilinear models. A few works worked on the recurrent neural network for caption generation. In [9], the system generate image descriptions using deep CNN(Convolutional Neural Network) and bag of words. Karpathy[10] produces a dense description of images using Region level CNN (RCNN) and bidirectional RNN. In [3], it is identified LSTM gives better performance for decoding operation. The system in [11] map the relationship between learned word embeddings and the LSTM hidden states. Authors generate captions using a deep Convolutional Neural Network (CNN) and two distinct LSTM networks for analyzing forward and backward direction of description. The approaches used in [12] systems are top-down and bottom-up.

## 3. Proposed System

Figure 1 represents the detailed architecture of the proposed system. In the proposed method, the system's primary goal is to enhance the automatic image description generation system's efficiency by generating meaningful sentences. The objects, attributes, actions, scenes, etc., are treated as image features. The sentence features are Noun, Adjective, verb, preposition, etc. The image captioning system maps the image features to sentence features. The essential tasks in this system are image parsing, sentence modelling, and surface realization. Preprocessing, feature extraction, and visual attention are the crucial steps in image parsing. Sentence modelling contains preprocessing and Text encoding. Caption generation, optimization, and Evaluation are the critical steps in surface realization.

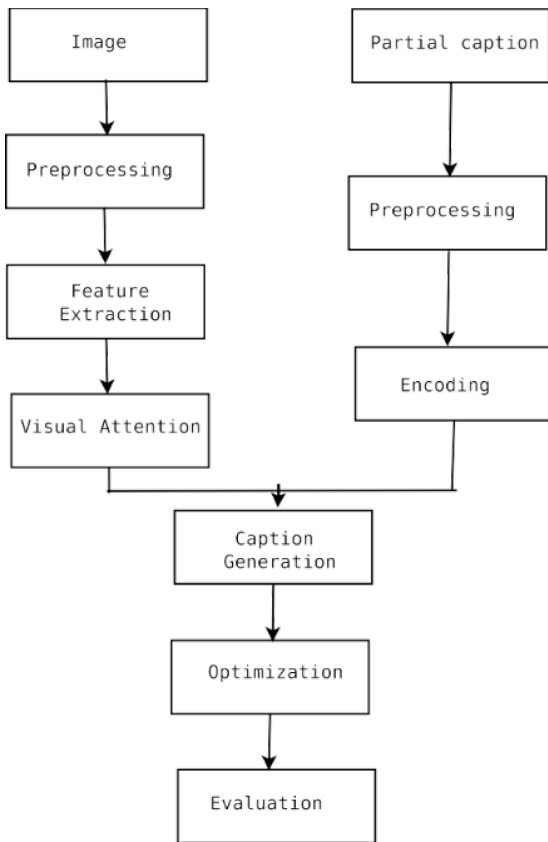


Figure 1: Methodology of Proposed System

Preprocessing is the first stage of the architecture, which brings the inputs image and partial caption into a normalized form that can be effectively dealt with by the systems (algorithms). The preprocessing is done on images and partial captions. In the current work, image preprocessing has been limited to image resizing and normalization. The essential steps in text preprocessing are vocabulary creation, word-to-index mapping, normalizing the length of the partial caption, word embedding vector creation, and one-hot encoding of output.

The Text encoding phase creates a hidden representation of the word embedding vector of a partial caption. Bidirectional LSTM is used for the text encoding phase.

The feature extraction phase recognizes the essential features from images needed for caption generation. The feature extraction is done in two ways: high-level feature extraction or keywords extraction and CNN feature extraction. From the experiments, CNN features are more suited for caption generation. Earlier feature extraction is done using local feature extraction techniques such as SIFT, SURF, etc., and global feature extraction techniques such as GIST, histogram, etc. Previous feature extraction techniques are time-consuming and not suited for the caption generation process. Deep learning-based feature extraction techniques give better performance on various tasks such as object classification, object detection, scene classification, etc. Various deep neural networks such as VGG, Residual Neural Network, and Densenet were experimented on the feature extraction process to determine the efficiency of the automatic description generation. The deep neural network with maximum performance is used in

the feature extraction process of the proposed methodology.

Visual attention is the process of finding a relevant part of the image suitable for the caption generation experimented. It is done on CNN features of images. A combination of spatial, channel-wise, and layer-wise attention was applied to improve the system's performance.

The caption generation phase produces the next word from the previously generated words in the description. Sequential models in deep learning are used in this phase. LSTM and Bidirectional LSTM have experimented with this process.

The optimization phase selects the good captions from the generated words. Beam search and game theoretic search are implemented for this process. Game-theoretic search outperformed beam search in description generation.

The evaluation model computes the system's performance using different evaluation metrics such as Accuracy, Loss, BLEU score, and ROUGE.

#### 4. Visual Attention

The visual attention model is based on a multi-attention system. The multi-attention module is made up of spatial attention, channel-wise attention, and layer-wise attention. The attention network used input as the feature maps from the second last layer of Densenet with a size of  $7 \times 7 \times 2208$ . An attention map and a score are the outcomes of the network. The captioning module takes the attention map and captions the image using an attentive region.

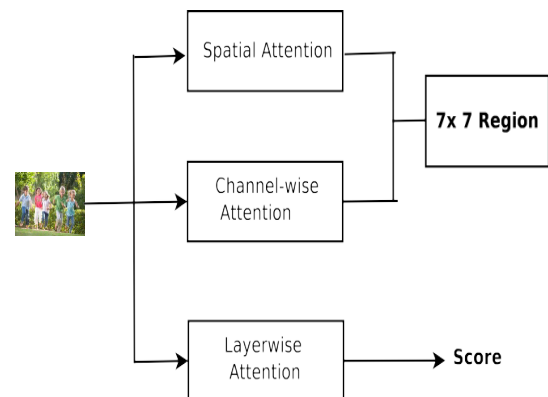


Figure 2: Visual Attention Architecture

The proposed architecture of the visual attention network is depicted in figure 2. Densenet is the convolutional neural network used for extracting the interest points of the whole image. The convolutional feature map is given to the attention network for getting a selected region and layer-wise scores. Different attention mechanisms used in the network are explained below.

##### 4.1. Spatial Attention

The usual image captioning systems use the global feature for generating descriptions. So, it is challenging to generate a correct caption for the image based on its regions. Only local regions are taken to get an accurate description. Some regions are more peculiar than other regions in an image. The critical regions in the picture are mainly helpful in producing better descriptions. In

spatial attention, more weights are given to the necessary region despite assigning equal weights for all regions in the image.

The network of spatial attention is shown in figure 3.

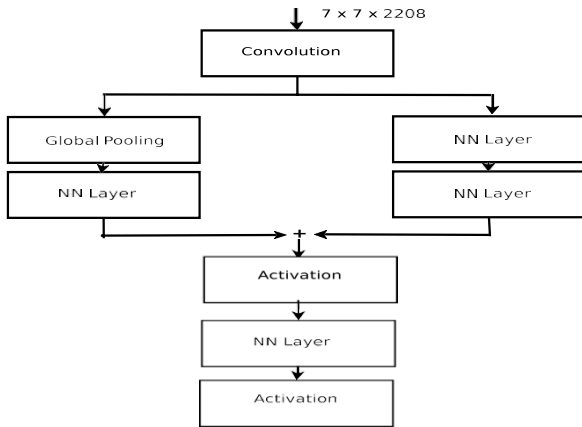


Figure 3: Spatial Attention

Let an image feature be  $F \in RW \times H \times C$ ,  $F$  is flattened as  $F_s = [f_1, f_2, \dots, f_n]$ , where  $f_i \in RC$ ,  $n = W * H$  and  $f_i$  represents a spatial region  $i$  through a vector  $f_i$ . The attention weight is calculated using a single layer fully connected neural network and a Softmax function over the  $W * H$  regions

$$S_a = \tanh((w_1 F_s + b_1) + (w_2 * F + b_2)) \quad (1)$$

$$S_w = \text{Softmax}(w_3 S_a + b_3) \quad (2)$$

where  $w_1, w_2, w_3$  are weight vectors and  $b_1, b_2$  and  $b_3$  are the bias values for the model.

#### 4.2. Channel-wise Attention

Colors and patterns are identified using CNN kernel functions. Some kernel functions are used to detect color information, and others are used for detecting the edges of the objects in the image. Channel-wise attention is a mechanism for choosing the channels dynamically; each channel of CNN features is obtained using the corresponding convolution kernel. The process of channel-wise attention is explained below.

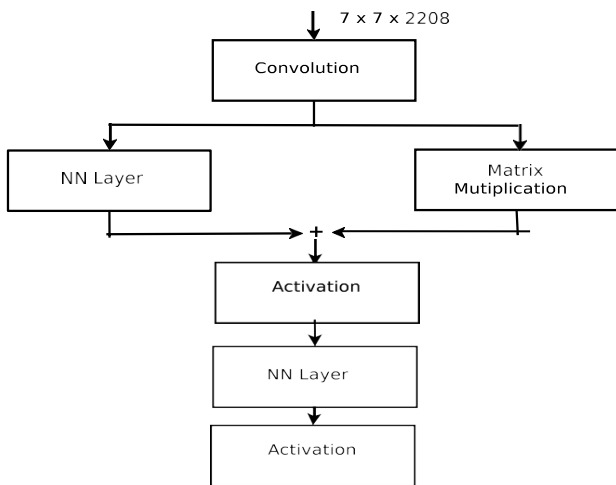


Figure 4: Channel-wise Attention

In this architecture, the image feature  $F \in RW \times H \times C$  is fed to the global average pooling function to produce the channel feature  $F_c = [F_1', F_2', \dots, F_c']$ ,  $F_c' \in Rc$  Where  $F_i'$  is the output of the global average pooling function on the feature of  $i$ th channel.

$$C_a = \tanh((w_1' F_c + b_1') + (w_2' * F)) \quad (3)$$

$$C_w = \text{Softmax}(w_3' C_a + b_3') \quad (4)$$

where  $w_1', w_2', w_3'$  are weight parameters and  $b_1'$  and  $b_3'$  are the bias values for the channel-wise attention model. The modelling of channel-wise attention is explained in figure 4.

#### 4.3. Layer-wise Attention

Different types of situations are handled by deep features in various levels. Layer-wise attention working is represented in the figure 5.

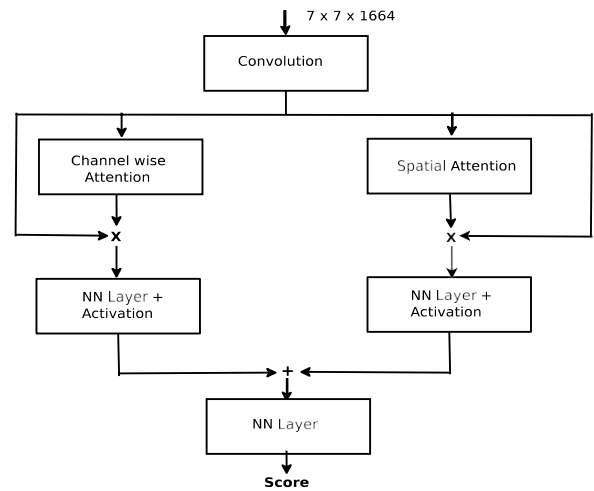


Figure 5: Layer-wise Attention

Given spatial weight  $S_w$ , channel-wise weight  $C_w$  and feature  $F$ , then

$$F_S = S_w * F \quad (5)$$

$$F_C = C_w * F \quad (6)$$

$$L_w = \text{ReLU}(w_{11} F_S + b_{11}) + \text{ReLU}(w_{21} F_C + b_{21}) \quad (7)$$

$$F_L = w_{31} L_w + b_{31} \quad (8)$$

where  $w_{11}, w_{21}, w_{31}$  are weight parameters and  $b_{11}, b_{21}$  and  $b_{31}$  are the bias values for the layer-wise attention model.

#### 4.4. Visual Attention Parameters

The attention model starts with a convolutional layer with a kernel of size  $1 \times 1$ , and the output of this layer is 512 channels. The width and height of spatial and channel-wise attention are 7. For spatial attention, the two fully connected network layers, matrix multiplication and activation functions, are integrated. The visual attention model is integrated into the caption generation system. So, the experiments are conducted for an image description generation system with visual attention.

## 5. Implementation Details

The framework was implemented using Keras, Tensorflow, and Python. Keras is a high-end deep learning package. The technology behind Keras is Tensorflow, which is a package for dataflow programming and machine learning.

### 5.1. Training Details

Image features are extracted using pre-trained Densenet model weights from ImageNet by the transfer learning mechanism. The language model used single-layer bidirectional LSTM with hidden size 256. Single-layer bidirectional LSTM, which has a hidden layer size of 1000, was employed in the caption model. The model is fitted to minimum validation loss at 50 epochs. Therefore, the model was finetuned with 50 iterations. In training, a random data generation method was employed in each iteration to limit computational resource usage. An NVIDIA Tesla K80 GPU is used for improving the training speed.

### 5.2. Optimization

The optimization function used was rmsprop. In rmsprop, the

learning rate for weight is divided by a running average of new gradients' magnitudes for that weight.


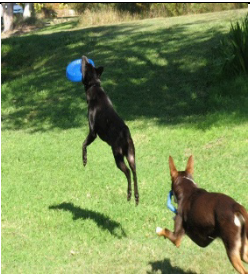
## 6. Experiments and Results


The performance of the model was analyzed using the BLEU[13] score. The BLEU score is an evaluation measure that compares the output description and the n-grams of the ground truth description of images. The Natural Language Toolkit (NLTK) package is used for computing the BLEU score. The experiment is done on two datasets, such as Flickr8k and ImageCLEF2019 datasets.

### 6.1. Flickr8k Dataset

The system is trained using a standard image captioning dataset Flickr8k[14]. The dataset is in the form of <image, caption>. The number of images in the dataset is 8000, and five captions are associated with each image. The training is done using 6000 images. The model is validated using 1000 images, and the testing is done on the remaining images. The correct image description results are shown in Table 1.



Table 1: Correct Results

Image	Ground Truth Captions	Generated Caption
	<ol style="list-style-type: none"> <li>1. A man crouch on a snowy peak.</li> <li>2. A man in green jacket stand in a deep snow at the base of a mountain.</li> <li>3. A man kneel in the snow.</li> <li>4. A man measure the depth of snow.</li> <li>5. A mountain hiker be dig steak into the thick snow.</li> </ol>	<p>Without VA+ Beam Search: A man with a stick in its mouth is standing on a snow covered field.</p> <p>Without VA +Game-theoretic search: A man with a stick is standing on a snow covered field.</p> <p>With VA+ Beam Search: A mountainer is standing in a snow covered field</p> <p>With VA +Game-theoretic search: A mountainer kneels in a snow covered field</p>
	<ol style="list-style-type: none"> <li>1. A dog with a Frisbee in front of a brown dog.</li> <li>2. A large black dog is catching a Frisbee while a large brown dog follows shortly after.</li> <li>3. Two dark colored dogs romp in the grass with blue Frisbee.</li> <li>4. Two dogs are catching blue Frisbees in grass.</li> <li>5. Two dogs are playing; one is catching a Frisbee.</li> </ol>	<p>Without VA+ Beam Search: A brown dog and a brown dog are running in a grassy field.</p> <p>Without VA +Game-theoretic search: Two brown dogs are running in a grassy field.</p> <p>With VA+ Beam Search: Two dogs are playing with a frisbee.</p> <p>With VA +Game-theoretic search: one brown dog and a dog are playing with a frisbee</p>

	<ol style="list-style-type: none"> <li>1. A man is sitting on the floor outside a door and his head on his chin.</li> <li>2. A man sits against a yellow wall wearing all black.</li> <li>3. A man wearing a dark blue hat sits on the ground and leans against a building.</li> <li>4. Man with black hat, coat and pants sitting next to the door of a building.</li> <li>5. The man in the black hat is sitting on the floor beside the green door.</li> </ol>	<p>Without VA+ Beam Search: A man in a blue jacket is sitting on a city street.</p> <p>Without VA +Game-theoretic search: A man in a blue jacket is sitting on a city street.</p> <p>With VA+ Beam Search: A man with a black hat is sitting near a door.</p> <p>With VA+ Game-theoretic search: A man with a black hat is sitting on a street near a door.</p>
---	---	---

The incorrect results are shown in Table 2.

Table 2: Incorrect results.

Image	Generated Caption
	A dog is playing with a ball
	A man with a stick in its mouth

The comparison of the model with various models is depicted in Table 3. The model was implemented with or without visual attention and caption optimization method as a beam search or game-theoretic search. Without visual attention and a game-theoretic search[15], the proposed model achieved a BLEU score of 69.96. The proposed model with visual attention and a game-theoretic search reached a BLEU score of 72.04, higher than all other models on the Flickr8k dataset given in Table 3. The results showed that the proposed model had a robust performance on the Flickr8k dataset.

Table 3: Comparison of the BLEU scores for different models.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
GoogleNIC [3]	63.0	41.0	27.0	-
Log bilinear[8]	65.6	42.4	27.7	17.7
Hard attention[16]	67.0	45.7	31.4	21.3
Soft attention [16]	67	44.8	29.9	19.5
Phi-LSTM [17]	67	44.8	29.9	19.5

Phi-LSTMv2 [18]	61.5	43.1	29.6	19.7
PhiLSTMv2(w.r ) [18]	62.7	49.4	30.7	20.8
Our Model (W/o Visual attention + Beam search)	67.2	55.05	44.42	40.61
Our Model (W/o Visual attention + Game theoretic Search)	69.96	56.3	46.45	42.95
Our Model (With Visual attention + Beam search)	71.2	57.2	46.97	43.21
Our Model (With Visual attention + Game theoretic Search)	72.04	58.0	47.23	43.95

The proposed system is also evaluated using ROUGE score. The ROUGE score is given in the table 4. ROUGE-1, ROUGE-2 and ROUGE-L scores are computed.

Table 4: ROUGE scores of the model

Score	Precision	Recall	F-score
ROUGE-1	58.1	56.21	57.13
ROUGE-1	58.32	56.34	57.31
ROUGE-2	44.25	42.82	43.52

GCorrect is an evaluation measure for measuring the grammatical accuracy of generated descriptions. GCorrect is the average of grammatical errors in the generated captions. It is defined by Equation (9).

$$GCorrect = \sum_{i=1}^n gerror_i / n \tag{9}$$

where gerror<sub>i</sub> is the number of grammatical errors for each

sentence and n is the number of sentences.

Grammatical errors in sentences were estimated using the Grammar-check package. The GCorrect of this framework is represented in Table 5 .

Table 5: GCorrect

Without Visual attention	0.040625
With Visual attention	0.023

## 6.2. ImageCLEF2019 challenge dataset[19]

The image caption pairs are extracted from PubMed Open Access. Seventy-two thousand one hundred eighty-seven radiology images are taken from the 6,031,814 image caption pairs after preprocessing. The number of images for training is 56,629 that for validation is 14,157 and for testing is 10000. Each label or symptom is mapped to a UMLS concept. The number of unique UMLS concepts is 5217. The examples of symptom UMLS concept mapping are depicted in table 6.

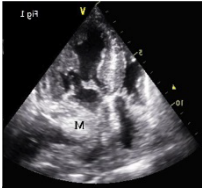
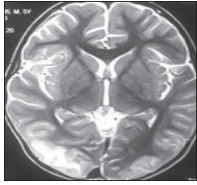

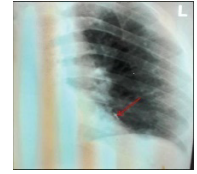
Table 6: Mapping of UMLS Concepts and Symptoms

UMLS Concept	Symptom
C0699612	osteogen
C0728713	sittings
C0376554	knowledge
C1262886	opg
C0152347	nucleus pulvinaris thalami
C0152344	capsula interna
C1550618	bronchial
C0152349	thalamus ventralis
C0520966	limb incoordination

Radiology images are resized to (224,224,3), and the intensity values of images are normalized between 0 and 1. Convolutional features of medical images are extracted using Densenet. Captions are preprocessed, and in the training set, each symptom is mapped to the concept. A vocabulary of concepts having a size 5217 is constructed. We tried our proposed image captioning model for this work.

The model is evaluated using the F1-score and mean BLEU score. Scikit-learn is used for calculating the F1-score. The default 'binary' averaging method is implemented. Some of the generated concepts with images are shown in table 7. The mean BLEU score and f1-score of the model for the medical image dataset are 25.30 and 20.56.

Table 7: ImageCLEF results

Sl. No	Image	Ground Truth Caption	Generated Caption
1		C0013516; C0203378; C0203379; C0183129; C0018792; C0221533; C0013524	C0013516; C0203378; C0203379; C0183129
2		C1552858; C0017067;  C0815275; C0015252; C1258666; C0007876; C0728940; C0007776; C0022655; C0184905	C1552858; C0015252; C0007876; C0007876
3		C0043299; C1548003; C1522577; C1962945	C0043299; C1548003; C1962945
4		C0700632; C1962945;  C1548003; C0179429; C0043299; C0817096; C0024109; C0796494	C1962945; C1548003; C1561542; C0043299

## 7. Conclusion

This paper mainly focuses on the generation of image descriptions using Deep learning methods and visual attention mechanisms. The automatic image description generation system uses encoder-decoder architecture. Different CNNs are considered for image feature extraction. Densenet gives better results for caption generation. The hybrid spatial, channel-wise, and layer-wise models are integrated into the image captioning system for producing high-quality descriptions. To optimize the words in the caption, a novel game-theoretic algorithm is introduced. Different language models are studied for generating descriptions, and

BLSTM is taken as the language model for our proposed system. An integrated framework for automatic image description generation was implemented. The model has experimented on both the general dataset Flickr8k and medical image dataset ImageCLEF2019 challenge dataset. The image and previously generated words are inputs to the integrated system. The system generated words sequentially. The system was evaluated using the BLEU score and ROUGE. The proposed method had a remarkable improvement over the state-of-the-art systems by five percentage. The grammatical correctness of the generated description was checked using a new evaluation measure called GCorrect.

## References

- [1] S. Kombrink, T. Mikolov, M. Karafiát, L. Burget, "Recurrent neural network based language modeling in meeting recognition," in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2011, doi:10.21437/interpeech.2011-720.
- [2] S. Hochreiter, J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, **9**(8), 1997, doi:10.1162/neco.1997.9.8.1735.
- [3] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, "Show and tell: A neural image caption generator," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015, doi:10.1109/CVPR.2015.7298935.
- [4] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A.C. Berg, T.L. Berg, "Baby talk: Understanding and generating simple image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**(12), 2013, doi:10.1109/TPAMI.2012.162.
- [5] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, H. Daumé, "Midge: Generating image descriptions from computer vision detections," in EACL 2012 - 13th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings, 2012.
- [6] A. Karpathy, A. Joulin, F.F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in Advances in Neural Information Processing Systems, 2014.
- [7] A. Farhadi, M. Hejrati, M.A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth, "Every picture tells a story: Generating sentences from images," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2010, doi:10.1007/978-3-642-15561-1\_2.
- [8] R. Kiros, R. Zemel, R. Salakhutdinov, "Multimodal Neural Language Models," *Proc NIPS Deep Learning ...*, 2013.
- [9] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2014, doi:10.1007/978-3-319-10593-2\_35.
- [10] A. Karpathy, L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**(4), 2017, doi:10.1109/TPAMI.2016.2598339.
- [11] M. Soh, "Learning CNN-LSTM Architectures for Image Caption Generation," *Nips*, (c), 2016.
- [12] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, "Image captioning with semantic attention," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, doi:10.1109/CVPR.2016.503.
- [13] K. R. S. W. T. Z. W. Papineni, "Bleu: A method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting on association for computational linguistics, 2002.
- [14] M. Hodosh, P. Young, J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, **47**, 2013, doi:10.1613/jair.3994.
- [15] S.R. Sreela, S.M. Idicula, "Dense model for automatic image description generation with game theoretic optimization," *Information (Switzerland)*, **10**(11), 2019, doi:10.3390/info10110354.
- [16] K. Xu, J.L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R.S. Zemel, Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in 32nd International Conference on Machine Learning, ICML 2015, 2015.
- [17] Y.H. Tan, C.S. Chan, "Phi-LSTM: A phrase-based hierarchical LSTM model for image captioning," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017, doi:10.1007/978-3-319-54193-8\_7.
- [18] Y.H. Tan, C.S. Chan, "Phrase-based image caption generator with hierarchical LSTM network," *Neurocomputing*, **333**, 2019, doi:10.1016/j.neucom.2018.12.026.
- [19] A.G.S. de H. and H.M. Obioma Pelka, Christoph M. Friedrich, "Overview of the ImageCLEFmed 2019 Concept Detection Task," in CEUR Workshop Proceedings (CEUR- WS.org), ISSN 1613-0073, <http://ceur-ws.org/Vol-2380/>, 2018.