AI Optics: Object recognition and caption generation for Blinds using Deep Learning Methodologies

Moksh Grover
Dept. of C.S.E.
HMR ITM
Delhi, India
mokshmg@gmail.com

Rajat Rathi
Dept. of C.S.E.
HMR ITM
Delhi, India
rajatrathi25@gmail.com

Chinkit Manchanda
Dept. of C.S.E.
HMR ITM
Delhi, India
chinkitm51@gmail.com

Kanishk Garg
Dept. of C.S.E.
HMR ITM
Delhi, India
kanishkgarg85@gmail.com

Ravinder Beniwal
Dept. of C.S.E.
HMR ITM
Delhi, India
ravin.beniwal29@gmail.com

2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) | 978-1-7281-8529-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCIS51004.2021.9397143

***Abstract*—With the exponential development in the field of artificial intelligence in recent years, many researchers have focused their attention towards the topic of image caption generation. With this topic being that of arduous task and interest people take it as a challenge to perform to excel in the field of AI. Automatic generation of neutral language descriptions or 'captions' according to the composition detected in an image, i.e., scene understanding is the main part of image caption generation which can be achieved by combining both natural language processing along with computer vision. In this paper, we tackle the task of generating captions by using the concepts of Deep Learning.**

Keywords— Artificial Intelligence, Deep Learning, RNN, CNN, LSTM.

I. INTRODUCTION

Millions of people around the world face major disability of visual impairment. Vision provides all the information needed for reading, body movement, mobility and its loss can severely affect an individual's professional and social advancement. It was reported by the World Health Organization (WHO) that out of 1.3 billion people that suffer from one or another form of visual impairment, 36 million suffer from complete blindness[1].

Problems are often faced by people with impaired vision or complete blindness once they are out of their familiarized environments. Corporeal development is one of the major issues for the people suffering from impaired vision [2]. They also are unable to recognize an object without physically feeling it and can'tsavor the beauty of the nature. Many assistive devices have been made commercially available for the visually impaired community of the society to help them read and recognize objects, enhancing their experience[3].

Various research works are still being done regarding the visually impaired community. Thorough analysis of a few papers has been done to understand the ongoing work and technology. A system that was to be worn in a shoe was proposed by K. Patil et al that contains ultrasonic sensors in all sides including vibration sensor, liquid detector and down step sensor. [4]. Other works included proposing android application for navigation and whether forecasting, news reading features using speech recognition and artificial intelligence. Electronic intelligent eye- a device was developed to implement range finder and camera for obstacle detection and navigation, the device also used solar panel for charging.

Computer Vision can be used to implement purposeful navigation and object detection for developing a technology for visual aid. Purposeful navigation refers to guided movement through free space to reach the desired location while prevention from hitting obstacles. [5]. The major challenge is to fast forward the results from the sensors to the processing algorithm and further to the accessible device.

In this paper we propose and end-to-end accessible solution to provide purposeful acknowledgement and guidance using object recognition and caption generation for enabling video to audio aid for the visually impaired community of the society. The aim of purposeful acknowledgement and guidance is to extract the range and direction of the obstacles within a finite and defined free space captured by the camera of the device. [6]. The object detection and recognition algorithm will be fed results to the caption generation algorithm for explaining the scene of the surrounding to the visually impaired in audio format. The same algorithm will also be responsible for providing guidance support to the blind. The paper specifically contributes the following:

1. A real time algorithm for mapping motion in free space using object detection.
2. Modified and explored version of general caption generator for feedback about the surrounding.
3. Capable system for providing guidance support through free space along with prevention from harmful and specific objects like fire, heavy traffic road, pointed ends, etc.

Automatic generation of a caption of an image is itself a big hurdle in artificial intelligence that involves connecting computer vision with natural language processing. [7]. But solution to this problem could prove to be a better understanding of the outside world for the visually impaired people. The task involves object detection and classification with high accuracy and accessibility with large of flexibility in inputs along with priorities to various situations. [8]. Previous studies have majorly focused on stitching together the solutions of the sub problems to form a larger solution and

have, therefore, failed in providing appropriate description to the image. [9]. In this paper, we propose a neural network based probabilistic model for the generation captions for images using a combination of recurrent neural network (RNN) and deep convolution neural network (DCNN) along with advanced statistical machine translation to obtain higher accuracy. [10].

The results from the caption generator are fed as the input to the guidance support system to map the purposeful motion of the user in free space using human body motion and mapping algorithm developed using modifications in CNN and statistical distance mapping done using python. The collective results of the proposed models are rendered to text to speech conversion algorithm in python to provide final output accessibility to the visually impaired.

II. Related Work

Computer vision technology has been used from a very long period of time for making description of visual data in the natural language [11,12]. Various types of systems have been developed for this purpose one such type of system is where structured formal language is joined with compound System. Type of such system are not very reliable as they are majorly hand created, have very few domains and are useful only on specific realm. [13].

Recently, object detection or image detection has gained a lot of popularity and interest of a large number of people. There are various kind of advances in the domain which aid in detecting natural language generation but are restricted in their outcome. Li et al [15] initiates with detection and combines them all-together to form a final outcome which consists of detected object and relationships. Similarly, Farhadi et al [14] utilized observations to produce a triplet of the image and changed them into text phrases with the help of template. Much greater model based on language parsing have been also utilized. The above methods have proved themself useful in various conditions but one issue that remains with them is that they are highly hand designed and rigid when used for text generations.

Various approaches have been also marked on the basis of problem of ranking descriptions [16,17,18]. Inthese kinds of method, the approach is to inserting text and images in the same vector space. Hence, when an image query is passed, those descriptions are fetched which lie near to image in the vector space.This approach cannot be used to describe new composition of objects, even if the separate object may have been noticed in training data.[19].

Latest image description can also be considered as dependent on language modelling which are created using recurrent neural network (RNN)[25,26,27]. The basic RNN model is a language model, its basic functioning is to arrest the probability of developing a string from the words generated so far. [20]. The RNN here is not only used to generated the next term of the string but also a set of image's features.So, here the RNN is a hybrid Model that functions and relies on both linguistic as well as visual features.

Kiros et al [21], proposed a multimodal log-bilinear model which was more inclined towards the features of the images. Later method was improved in such a way that to allow a natural process of generation and ranking. Donahue et al applied LSTMs (Long Short-Term Memory) to Video to generated video captions for the Video. LSTM is an artificial recurrent neural network is used in deep learning. IT has feedback connections. It is able to process single data as well as entire sequences of data [22,23].

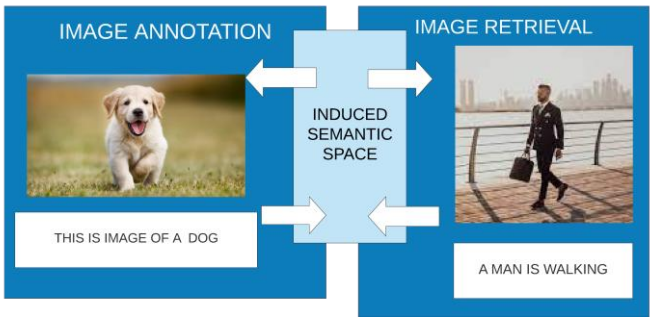


Fig. 1. Hybrid Model overview design

Fang et al described a three - step pipeline for generation by inculcating object detection. First process of their model was to learn detectors for various visual notion. Then a language trained model was applied to detector results, along with the image text embedding space.

In our work we combined image classification with deep convolution networks along with recurrent networks to developa single model that produces description (caption) of the images. The model is motivated by sequence generation, where instead of sentence an image is provided by CNN. A Latest work by Mao et al[24] used a NN for Same idea and outcome. Our used method is somewhat similar to Mao's approach with signifiable differences: - we have used more impactful RNN Model and the image is directly provided to the model directly. As a result, our system obtains a better result. Then we provided a multimodal embedding system space with RNN and LSTM that is used to remember text. Hence two separate pathways, i.e., one for image,while the other one for text to construe a joint embedding to produce speech outcome.

III. Background:Types of Architecture

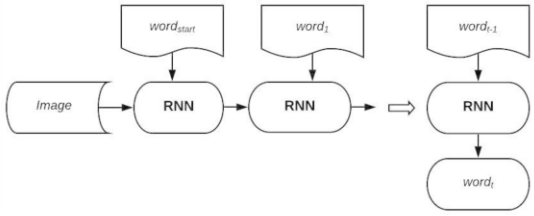
In the first section we constructed a prominent contradistinction amongst architectures that integrates rhetorical and image attributes in multimodal layer, and even those which inject the attributes of image straight onto the caption prefix encoding process.

We are also able to differentiate four rhetorical probabilities emerging from these architectures, as also depicted in the figures and briefed as following: -

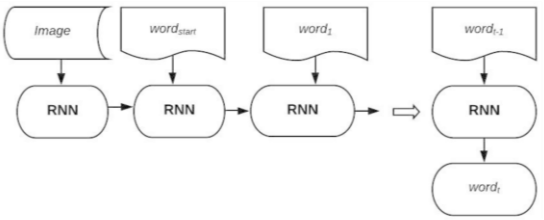
A. Init-Inject Technique

The initial state vector of the RNN is about to be an image vector (it can be also a vector that is extracted from image vector). It almost takes same size of the image vector as of the size of vector of RNN's hidden state. This is a static binding

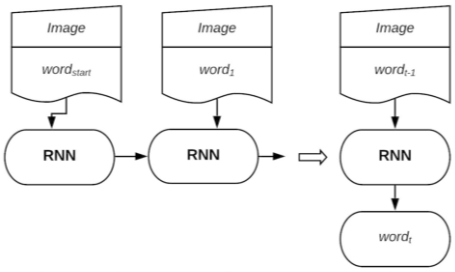
framework which also enables the descriptions of image to be altered by RNN.



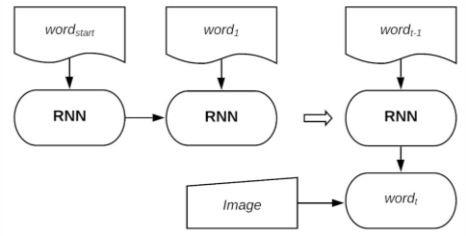
(a) Init-Inject: The vector of image is put to use as an hidden state vector for RNN.



(b) Pre-Inject: The vector of image is put to use as in the prefix as a first word.



(c) Par-inject: At each time step two inputs are Accepted by RNN. That are word and image.



(d) Merge: The prefix outside RNN and vector of image are merged together.

Fig. 2. Multiple techniques of constraining a neural language model alongside an image.

B. Pre-Inject Technique

RNN takes two inputs the first one isa vector that is extracted from image vector, i.e., image vector. The second input that is word vector comes into play later.

Therefore, in prefix the image vector is managed as the primary word. The magnitude of both the inputs that is image vector and word vectors must beequal. This is also a static binding framework and enables the image description to be altered by RNN.

C. Par-Inject Technique

The vector that is image vector or either extracted from image vector andthe word vector of caption prefix both simultaneously servers as inputs to the RNN in two ways or terminologies that are:-

a) Both the inputs are combined to form a single input (image vector is combined with word vector that is being forwarded to RNN.

b) RNN can also handle two discrete inputs.

As our previous possibilities both image vector and word vector of caption prefix needs to be the same size but, in this case, it is not required that every word vector has a corresponding image vector. Also, it is not required that each and every image vector and word vector must be similar. Therefor this is not a static binding architecture but rather is mixed unlike our previous case. Little bit of modification in also allowed while representation of the image. As it would be quite a task for RNN if every image that is fed to RNN is exactly the same as at its hidden state vector is refreshed with the same image every single time.

D. Merge Technique

The vector that is either derived from image or image vector is not exposed to the RNN at any instance. Rather than that the image is set forth in the language model following by the encoding of the prefix by the RNN. This is an example of a late binding architecture. [28].

We also don't require to alter the image at every time step during its representation. With these variations or possibilities, we are required to consider about a selection process of these above contributions.

IV. PROCEDURE

A. Prepare Photo And Text Data

For this following experiment we have used Flickr8K dataset which consists of two parts: 'Flickr8k_Dataset' which contains 8092 different type of photographs in '.jpg' or JPEG/JPG format and 'Flickr8k_text' which contains a number of text (.txt) files containing various sources of rawdescriptions for the given photographs.

Flickr8k dataset is separated into three sections:

a) For training purposes, we are provided with 6,000 images,

b) For testing purposes, we are provided with 1,000 images,

c) And for validating purposes we are provided with 1,000 images.

There are five different captions for each image. Using the VGG (Visual Geometry Group) class we load the VGG model

in Keras. we are curious about the photo's internal representation prior to classification is produced and not in the classification of images. From the pre-trained VGG-CNN we extract the 4096 element image feature vectors that are also available in the distributed datasets. During pre-processing these image vectors are normalized to unit length.

There is a unique identifier for each photograph which maps to a list of one or more textual description. these description texts need to be cleaned. These descriptions are easy to work with and already tokenized. Finally, we can summarize the size of vocabulary once we have cleaned the texts.

B. Develop Deep Learning Model

This section is divided into the following parts:

d) Loading the Datasets.

e) Defining the Caption Generation Model.

A. *Loading the Datasets*

All of the photograph along with their captions of the training dataset will be used to train the model. We can extract the photo identifiers using these file names. These identifiers are used to filter-out descriptions and photos for each set. A caption will be generated for a photograph that will be passed as an input for the model which will be created after a sequence of previously generated words are passed as an input, i.e., it will be generated one word at a time. In the final step we remove the 'startseq' and 'endseq' tokens and we have a base of our automatic caption generation model. For example, for "black dog is running in the water" as the input sequence we would have 8 input-output pairs for training of the model:

1	X1	X2(Text Sequence)	y(Word)
2	Photo	startseq	black
3	Photo	startseq, black,	dog
4	Photo	startseq, black, dog,	is
5	Photo	startseq, black, dog, is,	running
6	Photo	startseq, black, dog, is, running,	in
7	Photo	startseq, black, dog, is, running, in,	the
8	Photo	startseq, black, dog, is, running, in, the,	water
9	Photo	startseq, black, dog, is, running, in, the, water,	endseq

Fig. 3. Photo Identifiers being extracted using the file names.

The model will be trained in this fashion. Input text will be passed to a word embedded layer after they are encoded as integers. While we directly pass the photo features to the model in another section.

B. *Defining the Caption Generation model*

The model can be described in the following three parts:

a) Feature Extractor Layer: This is a 16-layer Visual Geometry Group model. This model is pre-trained on the ImageNet dataset. The photos are pre-processed with the VGG model after removing its output layer and the extracted features predicted by this model are used as input.

b) Sequence Processing Layer: The second layer is for handling the text input for which we use a word embedding layer, which is followed by a special kind of RNN layer called as Long Short-Term Memory (LSTM) layer.

c) Decoding Layer: We receive a fixed-layer vector as an output from both sequence processor and feature extractor. Finally, these both layers are integrated together and to make a concluding prediction, while these layers are processed by a Dense Layer.

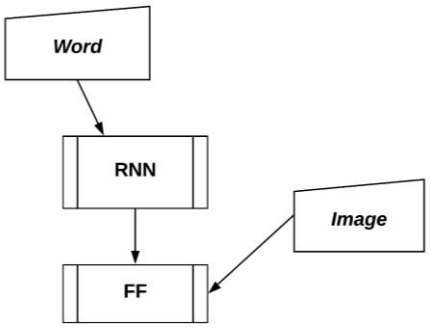


Fig. 4. Basic description of the model.

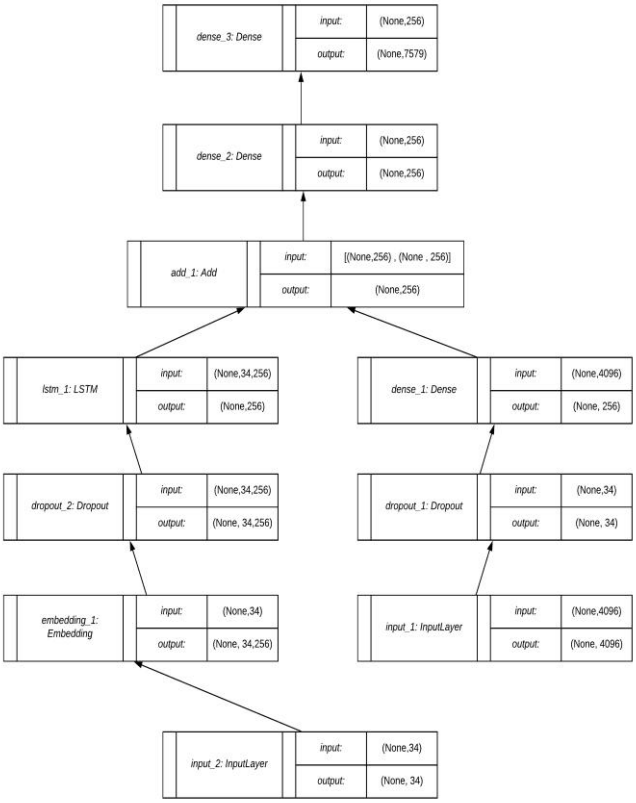


Fig. 5. Algorithm used in the model defined layer by layer.

The basic neural language model is made use of as a component of two dissimilar architectures in this experiment: inject architecture and merge architecture. Image vectors are interconnected sequentially with every one of the word vector in a caption in the Inject

Architecture. While in the merge architecture, image vector is connected sequentially with the final LSTM state.

An input of photographic features of a vector of 4,066 elements is expected at the Photo Feature Extractor Model, which are further processed by a dense layer. This layer compresses 4,066 elements used to represent the photograph to 256 elements.

While the Sequence Processor Model expects a predefined length of 34 words as an input sequence. This sequence is inputted into the 'Embedding Layer' in which the padded values are masked. After which a Long Short-Term Memory of 256 memory unit is attached.

A 256-element vector is produced by both of the input models. To reduce overfitting of the training dataset a 50% dropout regulation is applied on both of the input models, which results in fast model configuration and learning.

The vectors that are output of both of the input models are merged in the Decoder model using an addition operation. The product of addition operation is fed to a 256-neuron dense layer followed by another dense layer for the final output. This final layer makes arecursive prediction over the entire set of vocabulary output by using the next word in the sequence using the 'Softmax'.

Adam optimization algorithm was used to perform the training of this model with default parameters and 50 captions as the mini batch size. While sum cross-entropy was used as the cost function. An early stopping criterion was applied during the training. After each training epoch program measures the validation performance and once the performance on the validation data began to deteriorate training gets terminated.

V. RESULT

A. *Evaluate the model*

We have generated descriptions for all of the photographs present in the Flickr8K's testing dataset and then evaluate our model by evaluating these predictions with a standard cost function. We evaluate the actual and the generated descriptions by summarizing how resembling the predicted text is to the expected text. We accomplish this task by using the corpus BLEU score which are used during text translation, i.e., for the evaluation of generated translation text against a few reference translations. To evaluate the skills of any model we calculate the BLEU Scores for the 1, 2, 3 & 4 cumulative n-grams. We have some ball-park BLEU scores as a reference for skillful models when evaluated on the test dataset used in our experiment:

BLEU-1: 0.401 to 0.578.
BLEU-2: 0.176 to 0.390.
BLEU-3: 0.099 to 0.260.
BLEU-4: 0.059 to 0.170.

Fig. 6. BLEU Score range of a good model.

The following are our results of the BLEU Score of our model:

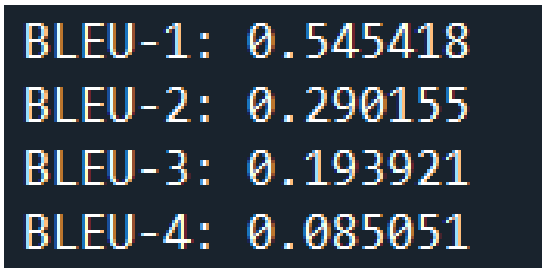


Fig. 7. BLEU Score values of model created.

We can see that the scores fit within expectancy range of aappropriate model on the specifiedquery and that too close to the top of the range.

B. *Generate Captions*

We load the photograph of which we want to generate the caption and extract the features from it. We achieved this by implementing the VGG-16 model after redefining our existing model else we can predict features using the VGG model and provide them to the existing model as the input.

Fig. 8. Input images provided to the model.

But how to generate a caption using our trained model? Initially we pass the 'startseq', i.e., the starting description token, generate one word and then recursively call the model again and again and pass the generated words as an input until maximum description length is reached or 'endseq', i.e., end sequence token is reached. In the final step we remove the 'startseq' and 'endseq' tokens and we have our caption for the photograph we passed for caption generation.

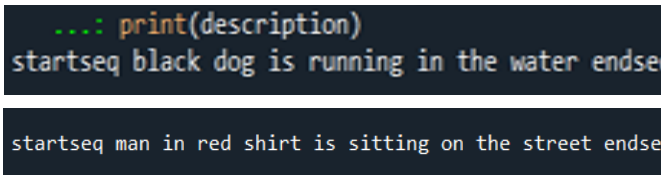


Fig. 9. Output captions generated by the model

VI. CONCLUSION

Various kinds of models are available today for video caption, image retrieval, image caption with their performance ability and the test results depicted that this system has greater performance. The modal basically focuses on the three important criteria:- first being on the generation of complete natural language sentences, second being on making the generated sentence semantically and grammarly correct and third making the caption consistent with the image.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Mar garet Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answer ing. In Proc. ICCV'15, pages 2425–2433, Santiago, Chile. IEEE.

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proce. Work shop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, vol ume 29, pages 65–72.

[3] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic Description Generation from Im ages: A Survey of Models, Datasets, and Evaluation Measures. Journal of Artificial Intelligence Research, 55:409–442.

[4] Xinlei Chen and C. Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image cap tion generation. In Proc. CVPR'15. Institute of Elec trical and Electronics Engineers (IEEE), jun.

[5] Manchanda, C., Rathi, R., & Sharma, N. (2019). Traffic Density Investigation & Road Accident Analysis in India using Deep Learning. 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). doi: 10.1109/icccis48478.2019.8974528

[6] Grover, M., Verma, B., Sharma, N., & Kaushik, I. (2019). Traffic control using V-2-V Based Method using Reinforcement Learning. 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). doi: 10.1109/icccis48478.2019.8974540

[7] Harjani, M., Grover, M., Sharma, N., & Kaushik, I. (2019). Analysis of Various Machine Learning Algorithm for Cardiac Pulse Prediction. *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. doi: 10.1109/icccis48478.2019.8974519

[8] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadar rama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term Recurrent Convolutional Networks for Visual Recog nition and Description. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). In stitute of Electrical and Electronics Engineers (IEEE), jun.

[9] Desmond Elliott and Frank Keller. 2013. Image De scription using Visual Dependency Representations. In Proc. EMNLP'13, pages 1292–1302, Seattle, WA. Association for Computational Linguistics.

[10] Manchanda, C., Sharma, N., Rathi, R., Bhushan, B., & Grover, M. (2020). Neoteric Security and Privacy Sanctuary Technologies in Smart Cities. *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*. doi:10.1109/csnt48778.2020.9115780

[11] R. Gerber and H.-H. Nagel. Knowledge representation forthe generation of quantified natural language descriptions ofvehicle traffic in image sequences. In ICIP. IEEE, 1996.

[12] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu. I2t:Image parsing to text description. Proceedings of the IEEE,98(8), 2010.

[13] Rustagi, A., Manchanda, C., & Sharma, N. (2020). IoE: A Boon & Threat to the Mankind. *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*. doi:10.1109/csnt48778.2020.9115748

[14] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young,C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picturetells a story: Generating sentences from images. InECCV,2010.

[15] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composingsimple image descriptions using web-scale n-grams. In Conference on Computational Natural Language Learning, 2011.

[16] M. Hodosh, P. Young, and J. Hockenmaier. Framing imagedescription as a ranking task: Data, models and evaluation metrics. JAIR, 47, 2013.

[17] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, andS. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In ECCV, 2014.

[18] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describingimages using 1 million captioned photographs. In NIPS, 2011.

[19] Sharma, N., Kaushik, I., Rathi, R., & Kumar, S. (2020). Evaluation of Accidental Death Records Using Hybrid Genetic Algorithm. SSRN Electronic Journal. doi: 10.2139/ssrn.3563084

[20] Rathi, R., Sharma, N., Manchanda, C., Bhushan, B., & Grover, M. (2020). Security Challenges & Controls in Cyber Physical System. 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT). doi:10.1109/csnt48778.2020.9115778

[21] R. Kiros and R. Z. R. Salakhutdinov. Multimodal neural languagemodels. In NIPS Deep Learning Workshop, 2013.

[22] Rustagi, A., Manchanda, C., Sharma, N., & Kaushik, I. (2020). Depression Anatomy Using Combinational Deep Neural Network. Advances in Intelligent Systems and Computing International Conference on Innovative Computing and Communications, 19-33. doi:10.1007/978-981-15-5148-2_3.

[23] Grover, M., Sharma, N., Bhushan, B., Kaushik, I., & Khamparia, A. (2020). 6 Malware Threat Analysis of IoT Devices Using Deep Learning Neural Network Methodologies. Security and Trust Issues in Internet of Things: Blockchain to the Rescue, 123.

[24] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille. Explainimages with multimodal recurrent neural networks. In arXiv:1410.1090,2014.

[25] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares,H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In EMNLP, 2014.

[26] D. Bahdanau, K. Cho, and Y. Bengio. Neural machinetranslation by jointly learning to align and translate. arXiv:1409.0473, 2014.

[27] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequencelearning with neural networks. In NIPS, 2014.

[28] Denkowski, Michael and Lavie, Alon. Meteor universal: Languagespecific translation evaluation for any target language. In Proceedings of the EACL 2014 Workshop on Statistical Machine Translation, 2014.