A Semantic Driven CNN – LSTM Architecture for Personalised Image Caption Generation

Abisha Anto Ignatious.L
Department of Information Technology
Anna University, MIT Campus
Chennai, India
jofinaanto@gmail.com

Jeevitha.S
Department of Information Technology
Anna University, MIT Campus
Chennai, India
sjeevitha887@gmail.com

Madhurambigai.M
Department of Information Technology
Anna University, MIT Campus
Chennai, India
madhurambigai@gmail.com

Hemalatha.M
Department of Information Technology
Anna University, MIT Campus
Chennai, India
hemalatham.ch@gmail.com

2019 11th International Conference on Advanced Computing (ICoAC) 978-1-7281-5286-8/20/$31.00 ©2020 IEEE 10.1109/ICoAC48765.2019.246867

Abstract* — Image Captioning is generating a human-readable textual description or a sentence about an image. The proposed semantic driven CNN-LSTM architecture comprises of the feature extraction process, semantic keywords extraction, facial recognition, and encoder-decoder LSTM networks. A pre-trained CNN is used to extract features from an image. A semantic keywords extraction module is used to identify the objects present in the image. The objects identified are labeled as the semantic tags present in the image. It increases the efficiency of captions in describing the objects and inclusion of these semantic labels in the captions. The LSTM based language model generates the captions by producing one word at a time. The facial recognition system identifies and recognizes the celebrity faces in the images, we have collected faces dataset which has facial images 232 celebrities. The instances of the person in the sentence were replaced with their names and personalized captions were generated. The Bilingual evaluation understudy (BLEU) and METEOR scores were generated to calculate the precision of generated captions.

***Keywords*—CNN, Image captioning, LSTM, semantics**

I. Introduction

The process of automatically generating the textual descriptions for an image is defined as image captioning or photo captioning. It describes the content of the image using text. It includes natural language processing for caption generation and computer vision for the prediction of attribute labels and other details of the instances and personalities. This process includes detection of objects, finding the relation between these objects, and prediction of their attributes. Every object has some unique features which describes the object in detail. The deep neural networks play a significant and promising role in identifying objects. The deep learning neural networks were able to learn from the already trained instances and predicts the output based on that knowledge. The level of abstraction in the layers of deep learning networks was increasing from layer to layer and output layer has more abstraction which predicts the labels and classes. The image captioning has recently gained an increasing interest in the field of information industries.

For the image captioning model, the system relies on Convolutional Neural Networks for visual identification and the language-based model, relies on the Recurrent Neural Networks. Convolutions are operated over the matrices which were built out of image pixel information. The output matrix detects the specific pattern of the input image and suggests the class labels using the values of the matrix. Usually, the CNN was used to extract the features from the image and it is linearly transformed into a feature vector. Then the LSTM networks is used as a language model to generate description.

In this paper we propose a personalized image captioning approach. All the existing image captioning approach does not identify the people in the images. Thus they provide a generic image captioning framework. The major contribution in this paper is as follows:

- Use of semantic labels to identify the objects present in the image
- Use of face recognition module to identify the famous personalities and celebrities.
- Integration of the semantic labels and face recognition modules to provide a personalized image captioning.

The proposed method is evaluated on Flickr30k and Faces dataset. The remaining part of the paper is organized as follows. Section II presents the literature survey. Section III discusses about the proposed approach. Section IV discusses about the experiments carried out and the results obtained. Section V provides the conclusion.

II. Related Work

The approach proposed in [6] automatically creates the description of an image using natural language processing. It involves both image processing and natural language processing. This paper discusses about the different models available for image captioning. They mainly focus on object recognition and machine translation for image captioning. They studied how these tasks have helped to improve the image captioning task. It finds the top-n matched images and its appropriate captions and these captions were taken as an output of the system.

The work in [5] propose the captioning task for large set of images with captions. The system tries to correlates the image features and keywords. It mainly discusses about the fact how the image features and the terms in captions were associated. The system used the blob-tokens generated using the G-means algorithm. This algorithm clusters the dataset, and blob-clusters were labels of the clusters in this system. A DT-RNN based model is proposed in [9] which uses dependency trees to embed sentences into vector space in order to retrieve images that are described buy those sentences. They compare the RNN and DT-RNN such that RNN uses the constituency trees which does not accurately represents the visually grounded meaning. DT-RNN abstracts from the details of the word order and syntactic expression.

The system proposed in [11] relies on the hardware requirements for processing the image captioning task. The system has an ability to run on low-end hardware of hand-held devices. It uses the encoder-decoder based implementation with significant modifications and optimizations which led to use it on hand-held devices. The method proposed in [3] present an approach to align image regions represented by a Convolutional Neural Network and sentence segments represented by a Bidirectional Recurrent Neural Network to learn a multimodal Recurrent Neural Network model to generate descriptions for image regions. It fused image features and visual attributes to conduct covert photo classification. The system proposed in [8] majorly concentrates on visual perception with great attention at the visual regions. The system also includes the scene-specific contexts that captures the important semantics information of an image. The proposed system works by extracting the visual regions of input image and simultaneously processing the scene extraction. It uses the LSTM networks for both finding the next visual focus should be and what the next word in the caption should be. The novelty lies in scene specific context because they improve the semantic information of the captions.

To reduce the impacts of noisy image estimations in available methods depends on image retrieval for image captioning [10] uses visual similarity to retrieve a set of captioned images for a query image. It employs computer vision algorithms to process an image and represent this image by using (objects, actions, spatial relationships) triplets. After that, they formulate image description as a tree-generating process based on the visual recognition results. Then, embed sentence fragments and image fragments into a common space for ranking sentences for a query image. They use dependency tree relations of a sentence as sentence fragments and use detection results of the Region Convolutional Neural Network method in an image as image fragments.

In the paper[4] a multimodal Convolutional Neural Network is proposed for matching images and sentences. It uses deep Canonical Correlation Analysis to match images and sentences. They use a deep Convolutional Neural Network to extract visual features from images and use a stacked network to extract textual features from Frequency-Inverse Document Frequency represented sentences. Template-based approaches have fixed templates with a number of blank slots to generate captions. In these approaches, different objects, attributes, actions are detected first and then the blank spaces in the templates are filled. A Conditional Random Field is adopted in [7] to infer the objects, attributes, and prepositions before filling in the gaps. Template-based methods can generate grammatically correct captions. However, templates are predefined and cannot generate variable-length captions.

The System proposed in [14] a facial action detection system. It works on the changing behaviour of facial muscle and analysing face expression. It gives 7.8 percentage more detection rate. This algorithm works with 46 action points comprising of the facial behaviour of human. A. The approach in [2] use detections to infer a triplet of scene elements which is converted to text using templates. The paper [13] used a face detection method, in this method set of rules are defined as a knowledge based method and it has rules on an input images first scanning is applied on face and all possible face can be found by scanning. Different rules are defined for face detection from image and at different level of rules are applied like at highest level set of rules are defined to define a human face and at lower level rules for facial feature are described. By averaging and sub sampling multilevel of hierarchy of images is created.

In [18] they have proposed a CNN-LSTM based approach to image captioning. Here they have extracted two features: features extracted using 2D-CNN and semantic features. They have used a LSTM in the language model. But they have not used a personalized framework for image captioning. In this paper a face detection module is used for generating a personalized image caption.

The approach in [1] proposes a robust facial action coding system. The Facial Action Coding System works on the changing behaviour of facial muscles and try to analyses facial expression. This algorithm gives 7.8% more average detection rate. The algorithm works with 46 Action Points (AP) comprising of the facial behaviour of human. A face detection technique as distribution based system and in this technique an object class learned using positive and negative examples [16]. Distribution based system have two components, multilayer perceptron classifier and distribution-based models for face non-face pattern. After normalizing and processing of each face example in a 19*19-pixel image, each image is vectored in 361-dimentional vector. Now using a modified K-means algorithm images are clustered into six faces and six non-face cluster. Each cluster is then represented as a multidimensional Gaussian Function with a mean image and covariance matrix.

III. PROPOSED WORK

A. System Architecture

The proposed system yields the output description of a given input images through the semantic feature extraction, face recognition developing a deep learning model. The Block Architecture for proposed image captioning system is depicted in Fig. 1.

An Encoder-Decoder model used to limit the length of the output description. Thus, the output has the descriptive fixed length sentence for a given or tested image. The image captioning is of divided into two modules, one is animage-based model which extracts the features from the image. The other module is a language model which translates the features and objects given by our image-based model to the natural sentence.

B. Semantic Keywords Extraction

The generated caption does not have more information about some attributes in the image. The Semantic key words Extraction from the description dataset will helps in improving the content quality of the captions. This process is similar to keywords extraction like TF-IDF instead of calculating the IF-TDF values.

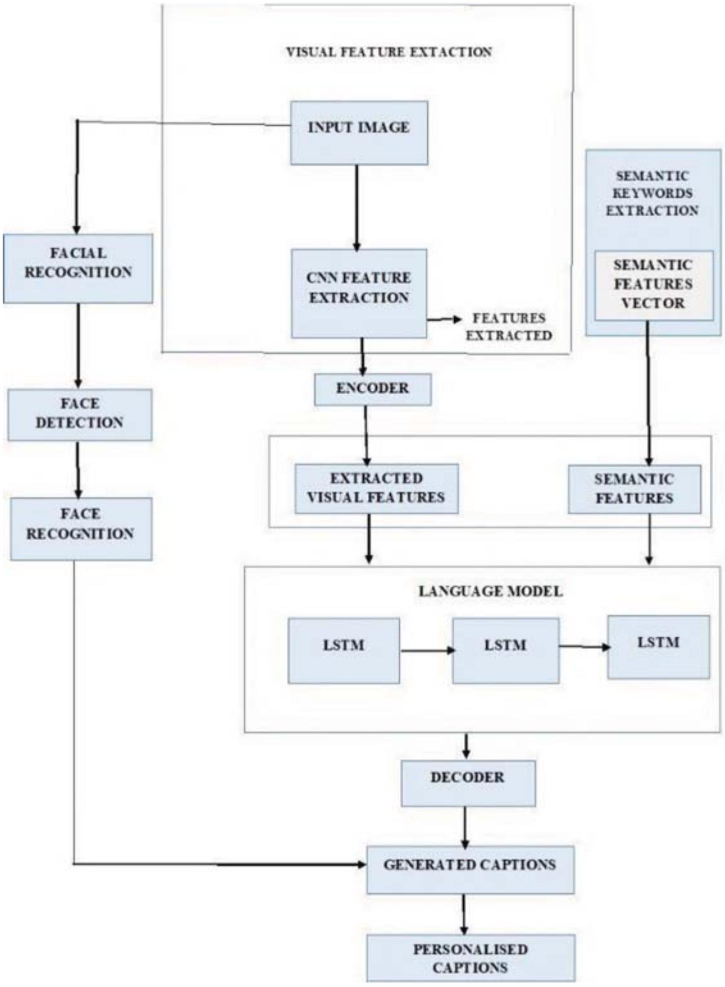


Fig.1. Personalized Image Captioning System Architecture

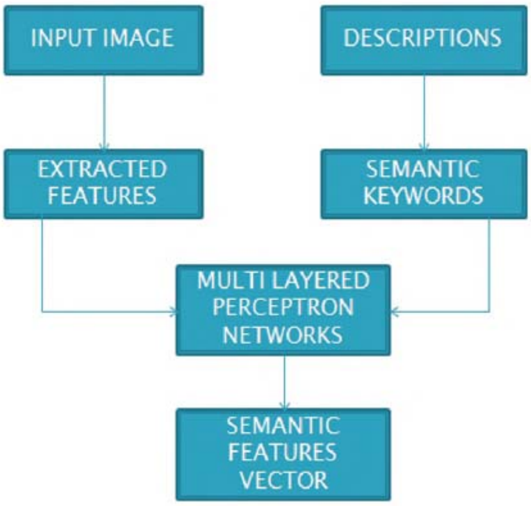


Fig.2. Semantic Keywords Extraction Module

The keywords were given as input to the multi–layer perceptron network which is shown in Fig.2. These network has hidden layers include dense layer, dropout layer, and soft – max layer which identifies the attributes in the image and helps in contributing the best captions with more information.

In every image we identify the significant objects present and name them as the semantic labels. These labels are provided as ground truth semantic labels to the semantic feature extraction module.

C. Steps in Semantic Keywords Extraction

Step 1: The extraction part involves cleaning the Flickr8k descriptions dataset by removing punctuations.

Step 2: Stemming and processing the description dataset.

Step 3: Spell checking the words using Python dictionary and convert it to vector.

Step 4: The Vector has been converted into a dictionary of unique keywords as key and their frequency as value.

Step 5: Selecting the top appearing 400 semantics from the Dictionary created in Step 4, an another vector is created.

Step 6: The vector in Step 5 is given as an input to the Multi-Layered Neural Network.

Fig.4 shows the sample images and the semantic features extracted from Flickr8k dataset using the proposed semantic feature extraction module. It is seen that the proposed semantic feature extraction module captures the important semantics present in the image. These semantics plays major role in generating the captions.

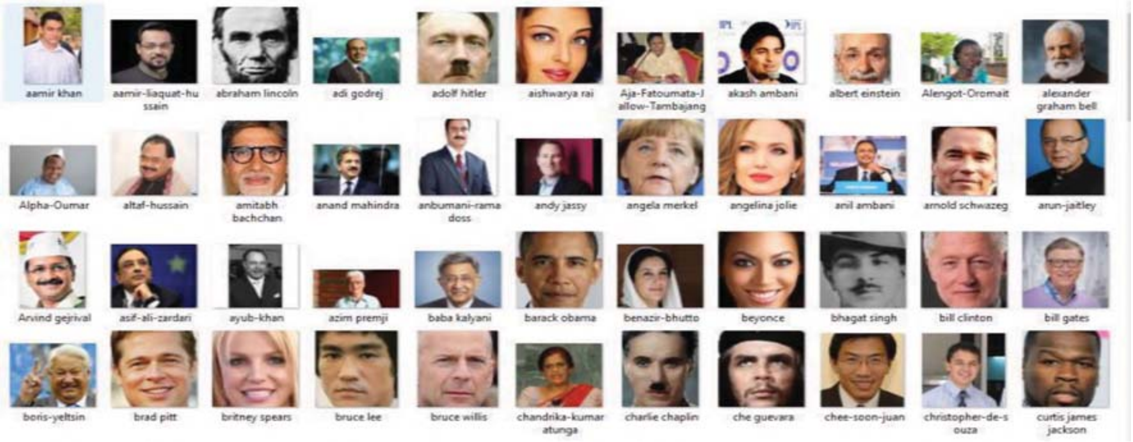


Fig.3. Faces Dataset

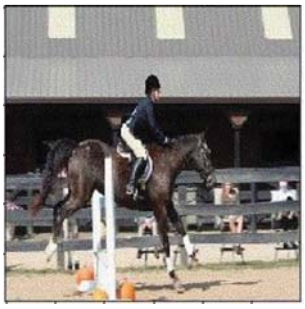
Image	Captions
	1: A jockey in a black horse jumps over a hurdle 2: a equestrian and a horse are jumping over an obstacle 3: a person wearing navy jacket and black hat jumping over a small partition on a horse 4: a show jumper is making a brown horse jump over a white fence 5: a woman on a horse jumps over an obstacle **Semantics: horse., woman jockey, jump, hat**

Fig. 4. Sample images and their semantics from Flickr8K dataset

D. Facial Recognition

Face recognition involves detecting faces in an image and to recognize who are in the image. There are two predominant approaches to the face recognition: Geometric (feature based) and photometric (view based). As researcher interest in face recognition continued, many different algorithms were developed, three of which have been well studied in face recognition literature. In our system, we use dlib and face recognition python libraries to recognize the faces. We have collected about 233 faces and recognition of the test sample images were based on this. The model has an accuracy of 99.38% on the Labelled Faces in the Wild benchmark. The dlib has 128 output feature vector which has values of eyes locations, chin height, cheeks thickness, nose width, etc.

One image for each face should be collected to recognize that person which is shown in Fig.3. The recognition process involves detecting faces, finding the facial encodings and comparing these encoding with test images. It matches the sample face encodings with dataset face encodings and returns the more probable one. The parameter tolerance is used to adjust which will uniquely identify the faces who has resemblance of other. The strict values will exactly identify the person.

E. Language Model

The language model for the proposed system was built using the RNN-LSTM networks because the experiments reveals that the RNN networks were suitable for text processing and generation process in deep neural networks. The features extracted from the upper layers and the semantic feature vector were given an input to the language model consisting of Recurrent neural layers. The sequence processor works with the help of the tokenizer class and generate sentence. The decoder generates the captions with the required length. Since, RNN layers has a memory unit which recognize which word should appear after the initial word and generates the sequence of words meaningfully which is shown in Fig.4.

The captions generated by the language model are not personalized which means that does not include the names of the famous personalities appearing in the image. Hence, the replacement methodology has been introduced to capture the names of the persons using face recognition process and replacing the formal pronouns with the appropriate name labels. The replacement of subjective terms for the images having more than one people involves concatenating the personalities name using a separator like commas, the final personalized captions were generated.

IV. EXPERIMENTAL RESULTS

The proposed system is developed using python, Tensorflow, Keras framework and the performance of the designed system is evaluated based on the Bilingual Evaluation Understudy (FLICKR8K) dataset [15]. The proposed approach shows an accuracy of 73.95% and it is tested upon the dataset which is a combination of a benchmark dataset and a real time dataset.

A. Dataset

The dataset for the image caption generation is collected from the FLICKR8K and it is pre-trained dataset specialized for captioning tasks.

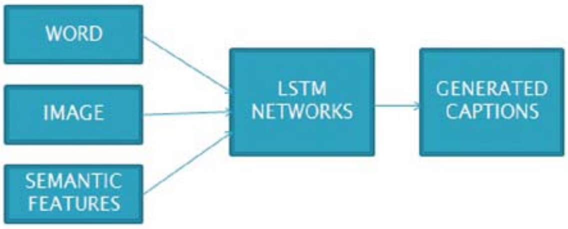
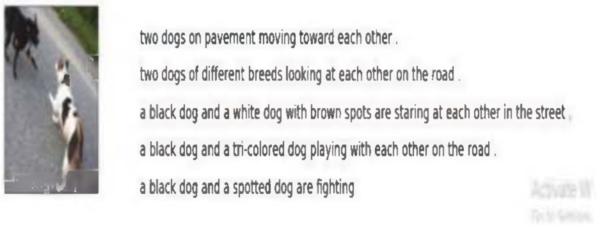


Fig.4. Language Model Architecture


two dogs on pavement moving toward each other .

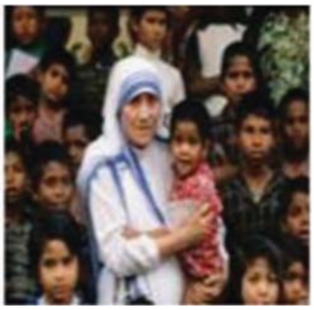
two dogs of different breeds looking at each other on the road .

a black dog and a white dog with brown spots are staring at each other in the street .

a black dog and a tri-colored dog playing with each other on the road .

a black dog and a spotted dog are fighting

Fig.5. Sample of Flickr8k Dataset



1.Mother Teresa is standing with children

2.Mother Teresa is lifting a child

3.Mother Teresa is wearing a white color dress

4.People are standing with mother teresa

Fig.6. Sample of Celebrity Dataset

For the personalization of the captions, the faces of various personalities from different fields was collected from random websites. The Flickr8K image dataset comprised of 8092 jpeg images collected from various websites. Each image has 5 captions recorded by humans which are admitted as ground truth which is shown in Fig.5 Each image has a unique ID.

The dataset includes images obtained from various website. It has images of various celebrities around the world. It majorly contains the cinema celebrities, popular business people and political giants. Five captions for each image is described because there is a great amount of variance that is possible in the captions that can be written to describe a single image. In the captions, the description is mostly based on the primary focused object in that image instead of explaining all the objects. The images were chosen according to their actions in the image and the captions has their dress attributes, their positions and their actions.

The celebrity dataset contains 362 images of popular politicians, business people, and on-screen celebrities accumulated from different web pages which is shown in Fig.6. Each image has a variable number of captions describing their attributes, actions, and gender. Each image has a unique ID. For the entity or face recognition of the celebrities, one image of each celebrity is collected. It has 233 faces in jpeg format which aids the face recognition system to compare the input faces with this dataset and to identify them.

B. Evaluation Metrics

BLEU (bilingual evaluation understudy) is scoring algorithm for evaluating the machine predicted text using automatic machine natural language systems. It works using the NLTK library of python. It has high precision with human results and it is inexpensive. The score values nearer to 1.0 indicate significant text quality. The value 0.0 implies the wrong and mismatched sentence. It works by matching the words count of n-gram candidate translation with n-gram in reference text, where grams represent the number of tokens taken for testing. There are variations of BLEU scores like sentence BLEU score, cumulative BLEU score, corpus BLEU score, individual n-gram scores and cumulative n-gram scores.

Several standard metrics such as BLEU-1,2,3,4 (precision-based) [15], METEOR (harmonic mean of precision and recall) [16],used for evaluating the image captioning system. All the four measures are evaluated in [14] for the image description problem and have identified that the METEOR score is better than BLEU and ROUGE-L. In this work, the performance of the proposed model is evaluated using BLEU and METEOR.

C. Comparison of BLEU Scores

The tested images and their predicted descriptions were quantified by BLEU scores. The BLEU score has a range of values between 0.0 – 1.0. The quality of the image and position of the person in that image also plays a major role for facial recognition. The tolerance value between 0.4-0.6 is most suited values. Flickr8k dataset caption generation without semantics gives BLUE score of 0.5 shown in Table 1 and Flickr8k caption generation with semantic gives BLEU score of 0.6.

The graphical representation explains the performance of the proposed system is shown in the Fig.7. It is seen that the performance is better for the web based celebrity caption dataset as the dataset includes more faces and provides personalized captions.

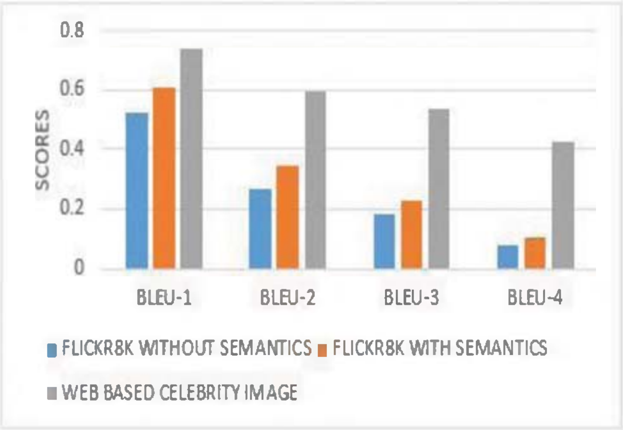


Fig.7. Performance graph of the proposed system

TABLE I. PERFROMANCE TABLE OF PROPOSED SYSTEM

Proposed methods	BLEU-3	BLEU-4	METEOR
Flickr8k caption generation without semantic features	0.18	0.08	17.5
Flickr8k caption generation with semantic features	0.33	0.23	18.8
Faces dataset - caption generation without semantic features	0.53	0.43	29.4
caption generation with semantic features	**0.70**	**0.55**	**30.2**

D. Implementation Details

The proposed model is implemented using Keras with Tensorflow backend. The ResNet152 models is used as feature extractor for the images. The ResNet152 model produces a 2048 number of features for every input image. The LSTM with 128 cells is used as language model. The Dataset is randomly divided into 60% for training, 20% for validation and 20% for testing. The model is training using Adam optimizer and the loss function used is categorical cross entropy function. The ReLu activation functions are used in the Dense Layers of the model.

E. Sample Input and Output

This section discusses the sample images and the captions generated by the proposed approach. Fig.8 and Fig.9 shows the sample image and their captions from Flickr8k dataset. The generated descriptions show that the image captioning system using the semantic features perform better than the image captioning system without using the semantic features.

FLICKR8K IMAGE SAMPLES

Image ID - ID of the image in the dataset.

Ground Truth – The descriptions suggested by humans for a specific image

Proposed system without semantic module – Description generated without inclusion of semantic features extraction module.

Proposed system with semantic module – Description generated with semantics extracted using the semantic feature extraction module.

SAMPLE 1:

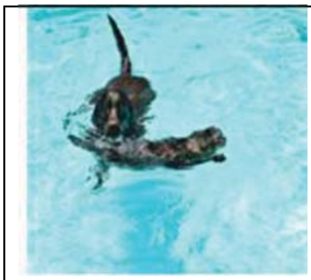
	**ID - 2599444370_9e40103027** **Ground Truth -** Two black dogs are swimming in a pool. **Proposed system without semantics module** – Dog is running through the water **Proposed system with semantic module** - The group of black dogs is swimming in water

Fig. 8. Sample image from flickr8k- 1

SAMPLE 2:

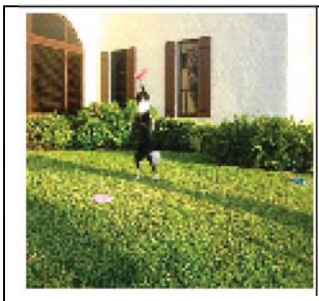
	**ID-15987659_b9eaa318dd3** **Ground Truth-**A black and whte dog is catching a Frisbee in the yard **Proposed system without semantics module** –Two children are playing in the grass. **Proposed system with semantic module-** The brown and white dog is playing in the grass.

Fig. 9. Sample image from flickr8k- 1

Fig.9 and Fig.10 shows the images and the captions from the web based celebrity dataset. It is seen that the proposed approach is able to detect the faces accurately and include the personalization in the captions. Here also the model with semantic features perform better compared to the model without semantic features.

WEB BASED CELEBRITY DATASET SAMPLES

SAMPLE 3

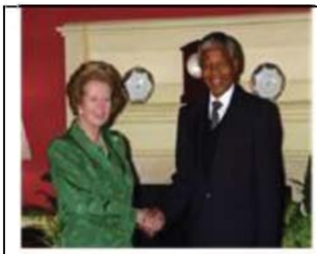
	ID – 3100 **Ground Truth** – Margaret Thatcher is standing and shaking hands with Nelson Mandela **Proposed Personalised caption generator** – Margaret Thatcher is handshaking with Nelson Mandela

Fig. 9. Sample image from web based celebrity dataset- 1

SAMPLE 4

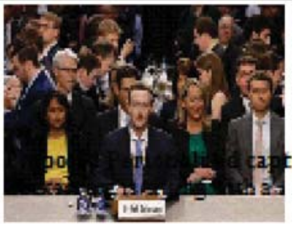
	**ID – Real world image** **Ground Truth** - Mark Zuckerberg is wearing black dress and sitting with a group of people **Proposed Personalised caption generator-** Mark Zuckerberg is sitting with a group of people

Fig. 10. Sample image from web based celebrity dataset - 2

V. CONCLUSION

The image captioning has gained a lot of attention in artificial intelligence scope and nowadays many applications were based on this technology. The proposed system is a deep learning based LSTM and CNN places in the image and uses it to generate a sequence of words that describe content of the image. The inclusion of semantic classifier in feature detection process makes the caption to explain the actions of the objects based on the scene in that image. The experimental results show that the RNN networks play a crucial role in determining the sequence of words in the sentence. The main purpose of the proposed system is to enhance the captions through semantic extraction process. The language model designed using LSTM layers performs well in generating sentences. The Facial Recognition was involved to generate the captions in a personalized manner which gains a performance of about 93.8%. The captions generated with extraction of semantic features has achieved the better performance compared to caption generation without semantic features.

ACKNOWLEDGMENT

This work is funded by Centre for Technology Development and Transfer, Anna University, Chennai-25, Tamil Nadu under Student Innovation Project Scheme.

REFERENCES

[1] P. Ekman, W.V. Friesen, "Facial action coding system: investigator's guide", Consulting Psychologists Press, Palo Alto, CA, 1978

[2] A. Farhadi, M. Hejrati, M.A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth, "Every picture tells a story: Generating sentences from images", ECCV, 2010.

[3] M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R.Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification", IEEE Conference, 2015

[4] S. Ji, W. Xu, M. Yang, and K. Yu, "Convolutional Neural Networks for human action recognition", IEEE Transactions, 2013

[5] Jia-Yu Pan, Hyung-Jeong Yang, P. Duygulu, C. Faloutsos, "Automatic image captioning", IEEE International Conference, 2004

[6] Kun Fu, Junqi Jin, Runpeng Cui, Fei Sha, Changshui Zhang, "Aligning Where to See and What to Tell: Image Captioning with Region-Based Attention and Scene-Specific Contexts", IEEE Transactions, Jun 2015

[7] G. Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg, "Understanding and generating image descriptions",2011, In Proceedings of the 24th CVPR

[8] Parth Shah, Vishvajit Bakrola, Supriya Pati, "Image captioning using deep neural architectures", IEEE Conference, 2017

[9] Pranay Mathur, ,Aman Gill, Aayush Yadav, Anurag Mishra and Nand Kumar Bansode,"Camera2Caption: A Real-Time Image Caption Generator", IEEE International Conference, 2017

[10] M. H. Peter Young., A. Lai, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions", IEEE transactions, 2014

[11] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences", IEEE transactions, 2014

[12] K. K. Sung, "Learning and example selection for object and pattern detection", IEEE Transaction ,2017

[13] G. Yang, T. S. Huang," Human face detection in a complex background. Pattern recognition", 2015, vol. 27, no. 1, pp. 53-63

[14] Hodosh, Micah, Peter Young, and Julia Hockenmaier. "Framing image description as a ranking task: Data, models and evaluation metrics." Journal of Artificial Intelligence Research 47 (2013): 853-899.

[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318., 2002.

[16] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 2015.

[17] Michael Denkowski, Alon Lavie, "Meteor Universal: Language Specific Translation Evaluation for Any Target Language," Proceedings of the ninth workshop on statistical machine translation, 376-380, 2014.

[18] Aravindkumar S., Varalakshmi P., Hemalatha M. "Generation of Image Caption Using CNN-LSTM Based Approach." Intelligent Systems Design and Applications. ISDA 2018. Advances in Intelligent Systems and Computing, vol 940. Springer, Cham.