iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 Issue: I Month of publication: January 2022

DOI: https://doi.org/10.22214/ijraset.2022.40057

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue I Jan 2022- Available at www.ijraset.com

Image Captioning with Face Recognition using Transformers

Mohd Wasiuddin Junaid[1], W. Wilfred Godfrey[2], Jeevaraj .S[3]

***Abstract:** The process of generating text from images is called Image Captioning. It not only requires the recognition of the object and the scene but the ability to analyze the state and identify the relationship among these objects. Therefore image captioning integrates the field of computer vision and natural language processing. Thus we introduces a novel image captioning model which is capable of recognizing human faces in an given image using transformer model. The proposed Faster R-CNN-Transformer model architecture comprises of feature extraction from images, extraction of semantic keywords from captions, and encoder-decoder transformers. Faster-RCNN is implemented for face recognition and features are extracted from images using InceptionV3 . The model aims to identify and recognizes the known faces in the images. The Faster R-CNN module creates the bounding box across the face which helps in better interpretation of an image and caption. The dataset used in this model has images with celebrity faces and caption with celebrity names included within itself, respectively has in total 232 celebrities. Due to small size of dataset, we have augmented images and added 100 images with their corresponding captions to increase the size of vocabulary for our model. The BLEU and METEOR scores were generated to evaluate the accuracy/quality of generated captions.*

***Keywords:** Image Captioning, Faster R-CNN , Transformers, Bleu score, Meteor score.*

I. INTRODUCTION

A picture is worth of thousand words, and truly large amount of information is conveyed at a single glance on an image. We humans usually interact with fellow human through spoken or written languages. Humans have the ability to generate different captions for given identical images. Image captioning is one of the emerging topic in the field of artificial intelligence, If we can automate this task with the help of machines, it will be beneficial in various tasks such as automatic medical report generations, aid to blind people(image to captions, then captions to voice), self driving cars, and better CCTV monitoring in case of tragedy etc.

Thus, this work aims to contribute to image captioning with face recognition task. This computer vision system aims to recognize the known person in an image, localize and describe salient regions in images in natural language. The Image captioning task generalizes object detection when the descriptions consist of a single word. Given a set of images and prior knowledge about the content, find the caption for the entire image. Thus it provides a generic image captioning framework which is capable to recognize the faces, making image captioning more relatable and relevant. We also incorporated Faster R-CNN module for face recognition and to draw abounding box around the face which helps in better interpretation of an image and caption. In the current scenario generating captions from an image is a very challenging task for machines. Generating human-readable captions from an i-age is still a challenging and exciting task that requires a concise knowledge of natural language processing and computer vision that can identify and correlate objects in an image. Despite being highly complex task, many works have been done to decode image captioning problem. Recent developments in deep neural networks and the availability of large classification datasets like ImageNet have paved the path to not only resolve this problem but also helps in achieving better results in terms of generating quality captions with the help of CNN as encoder and RNN [3,5], LSTM [1]. and Transformers [4] as decoder. Most of the conventional image captioning are not capable of identifying the people in the images. Thus, we introduce a model based on transformers for image captioning with face recognition.

II. LITERATURE REVIEW

Most of the conventional image captioning models yield to output just captions for the given images without recognizing faces. Thus this work presents novel image captioning with face recognition using transformers.

Initially, image captioning was attempted to briefly describe the images taken under extremely cramped conditions and were not focused on generating text from real-life images. From Early stages to recent times, various methodologies have been developed for an image-to-text generation. One of those methods is retrieval-based image captioning. In real it does not actually generates text,rather for an input images it retrieve the sentence from already given pool of sentences or composed from set of retrieve sentences. The works includes Farhadi et al model [9].

The authors establishes a links between images and sentences using Markov Random Field, and use Lin similarity measure to calculate the semantic distance between this image and each existing sentence. The sentence closest to the given input image is taken as its caption. Hodosh et al.[10] considered image captioning as a ranking task. In this model authors tried to maximize the correlation between image and sentence during training phase. Then cosine similarity between images and captions are calculated and top ranked sentences are considered for an given image. Since the generated captions are already provided is not capable of generating captions for novel scenes and limited to images in the dataset.

The other method includes template based image captioning. In this method,the structure of the sentence is predefined. Then the detected objects are connected to predefined sentence template to generate the text from an image.Kojima et al.[7] used a conceptual hierarchy of actions, case structures, and verb patterns to generate natural language to explain human activity in an office environment. Hede et al[8] used a dictionary from object and language template for describing object images in Chaotic background.Because these methods accomplish the image captioning task by relying on hard-coded language structures,the disadvantage of this method is that they are not flexible enough. As a result,expressiveness of generated descriptions by these methods is, to a large extent,limited.

Due to the complete dependence on given pool of sentences on and lack of actual learning for creating text by itself in retrieval based models and dependence on hard-coded language structures in template based model. Thus, these models lacks in creativity. Therefore these models are far from being used to describe the images in recent times.Recent works on image captioning were based on deep learning techniques which uses pre-trained CNN as an encoder, and the last hidden layer of the CNN is given as input to RNN i.e decoder. Mao et al.[12] has implemented image captioning with CNN as encoder and RNN has decoder. RNN language model is employed to calculate the probability for each word in sentence conditioned on image. The problem with RNN is that,it finds difficulty in learning long-term dependencies which means it may not produce better results with increase in length of the caption. Hence LSTM were introduced to resolve this problem. Vinyals et al.[11] proposes convolution neural network as encoder to extract features from images and LSTM to decode the extracted features in the sentence. The LSTM also works in the similar way as of RNN. In the current study, similar encoder–decoder-based transformer architecture[2] as decoder is implemented instead of RNN or LSTM. The proposed model is based on multi-stacked attention transformer model.It is different from the conventional sequence to sequence model because it does not use an RNN. Transformers have been implemented as a decoder in the proposed model. Therefore, RNN as a decoder is absent in the proposed model.The drawback with RNN based models [6] is that, Parallelization is not possible because current hidden step is computed based on previous step therefore,it has to wait for previous state. Hence, parallelization is not possible. Thus it creates a significant problem with GPUs. Due to the sequential nature of RNN,GPUs are not compatible and are not efficient as they have to wait for data to become available. The RNN continues to be challenging because, it is incapable of handling long-range dependencies. There is also very limited work done on image captioning model which is capable for recognizing faces in an image. But there is one previous work on similar problem[1]. The authors in this paper [1] work on image captioning which also detects the faces but by using CNN-LSTM architecture. They had implemented through ResNet152 pre-trained CNN model and LSTM is used as language model. We in this work introduces a transformer neural networks which had proved to have better accuracy than LSTM. The approach in [1]comprises conventional image captioning with separate face recognition using dlib library. We aims for an different approach where Faster R-CNN module[64] is implemented for face-recognition. The idea behind using Faster-RCNN instead off ace-recognition library is that, Since conventional image captioning models and morever Faster R-CNN requires CNN as a base rather than using face-recognition library and CNN separately [1]. However, Faster R-CNN aids in better visualization by drawing bounding box around the faces helps in better interpretation in image where more than one person and also when celebrity is unknown to user present in the image.

III.PROPOSED WORK

A. System Architecture

The proposed model produces the caption for a given images by implementing image augmentation, feature extraction from images, and processing the captions. The architecture for proposed image captioning system is represented in Fig. 1

Firstly images are augmented using various techniques to enhance training the of model. A Faster R-CNN-Transformer model is used to limit the length of the output description as there is a possibility of generating description of infinite words.Thus, the threshold of maximum 30 words in a description have been set for a given image. The image captioning is broadly categorized into two modules,one feature extraction from the images with help of pretrained CNN-InceptionV3, and secondly transformer translates the features and objects in an image to a natural sentence. The noun in the generated caption is replaced by the name of the celebrity in the image recognized by the faster R-CNN.

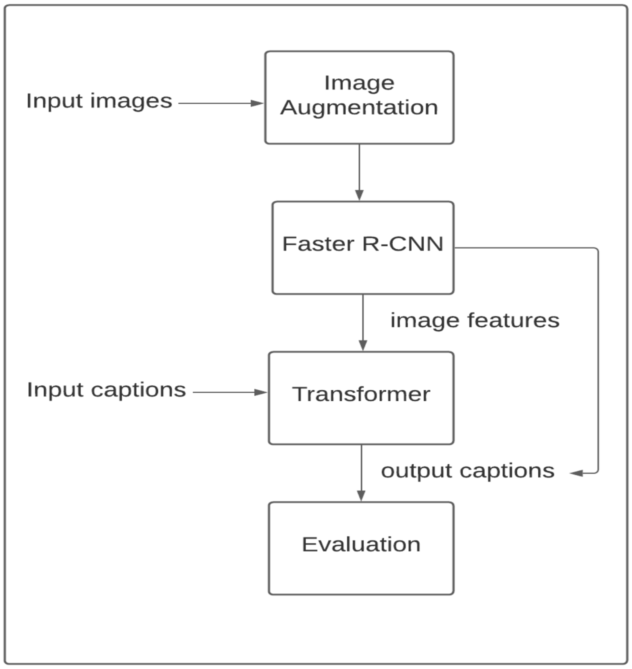


Fig. 1 proposed model

B. Data Augmentation

Due to inadequate data, images are augmented using following techniques [6].

1) Horizontal flip.

2) Brightness.

3) Zoom-out.

Also,100 random images and 500 corresponding captions from public dataset called flickr were merged to the dataset to increase the size of the vocabulary to improve the semantics of generated captions. The total images and size of vocabulary (all the unique words in the dataset) obtained after augmentation are 1450 and 800 respectively.

C. Faster R-CNN

Faster -RCNN comprises of CNN and Region proposal network(RPN)[13]. It can be explained in a following way. The input is given to pretrained CNN and features are extracted.The structure of RPN is a small convolutional neural network. We input the feature map area and output two layer Classification Layer and regression layer extracted with 1 x 1 convolution kernel. The classification is binary classification to find out whether object is present or not. the regression layer gives out the coordinate of bounding boxes of our class and this region can be refered as region of interest (RoI). Now after RPN pooling is applied to RoI to resize them in a uniform size. Later it is passed through fully connected layers to find the name of object in RoI region.

The role of CNN is of encoder which extract the features from images. This model is implemented using pretrained InceptionV3[17] on Imagenet dataset and fine tuned with the augmented image dataset [6] to extract features from images. Hence softmax layer is removed from the model. All the images were re-sized to the uniform size, before giving as input to CNN. The features are extracted and stored as numpy files (.npy).Then the extracted features are mapped to respective images names and fed tot he encoder of transformer [2] and RPN block of Faster RCNN. The final layer of Faster R-CNN detect the celebrity and bounding box co-ordinates for better interpretation of image and caption.

D. Transformer

The role of Transformer model with a stacked attention mechanism is of decoder. Here in place of single attention, multiple stacked attentions for a single input is employed in the model [5]. Along with the attentions, it uses residual connection and normalization layer to make optimization faster and easier. To maintain the position of the input, the current model also uses an explicit positional encoding as depicted in Fig. 2.

The encoder of transformer takes the image features as input and applies attention mechanism on those image features. Then it is passed to the decoder of the transformer. The tokenized captions are assigned with certain weights based on the semantic meaning and predefined features in embedding layer. For example, a man and a boy are assigned similar weights as they signify similar meaning. Whereas pen and pigeon are assigned very different weights since they are not related. Since, transformer takes the parallel input. It becomes extremely important to maintain the order of each word in captions. The task to keep track of the words in captions is taken care by positional encoding layer.

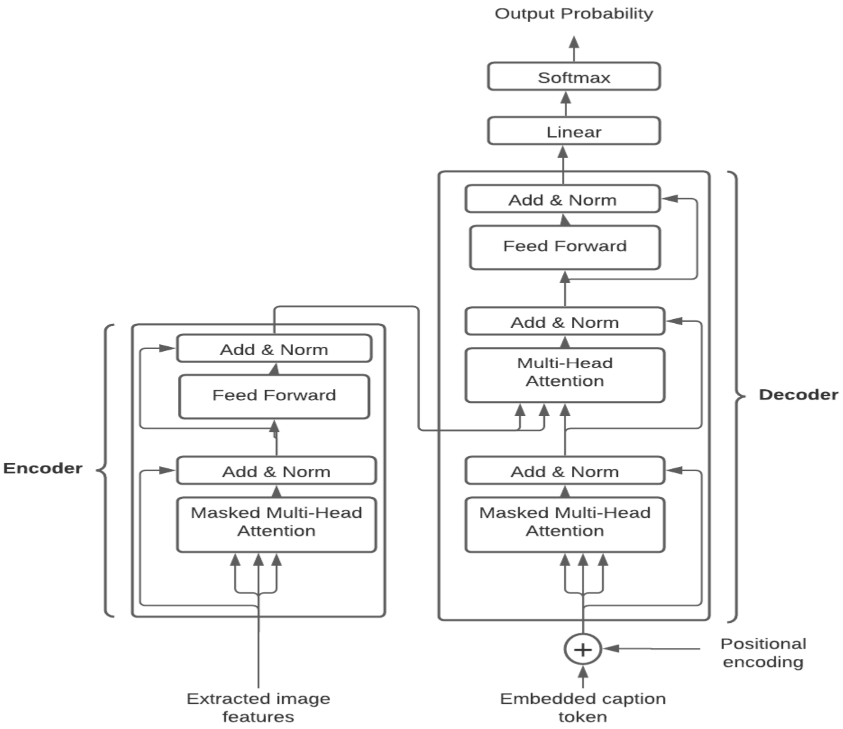


Fig.2 Transformer

Decoder takes the attention weights from encoder of the transformer for training. The decoder of the transformer also takes the pre-processed captions and train the model with corresponding images features. The output from this layer is passed to fully connected layers to predict the captions. Scaled Dot-Product Attention is used to calculate the attention in trans-former. It consists of: keys K, values V, and queries Q. This is computed as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}(Q*K^T / \sqrt{d_k}) \qquad (1)$$

Here Q is the query matrix and K and V are the matrices of keys and values respectively. d_k is the dimensionality of queries or keys. This attention mechanism is space-efficient as well as faster than others. It can be implemented using an optimized matrix multiplication method.The Idea behind multiplication of Q, K (transpose of K is taken because it is a matrix) and divided by square root of dimensions of K can be analogous to cosine similarity, and softmax functions normalizes the values between 0 and 1 which simplifies computation

Multi Head Self Attention Mechanism is a module for attention mechanisms which runs through an attention mechanism several times in parallel. The independent attention outputs are then concatenated and linearly transformed into the expected dimension. Intuitively, multiple attention heads allows for at-tending to parts of the sequence differently (e.g. longer-term dependencies versus shorter-term dependencies).

$$h = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \qquad (2)$$

$$H = \text{Concat}(h_1, h_2, h_3, .., h_n) \qquad (3)$$

$$O = HW_h \qquad (4)$$

Above W are all learnable parameter matrices. Scaled dot-product attention is most commonly used in this module, although in principle it can be swapped out for other types of attention mechanism.

In the current model, Transformer within itself is broadly categorized in to two blocks. one is encoder and other is decoder.Both of them consists of similar attentions. The only difference between encoder and decoder is masked multi-head decoder self-attention (the first sub-layer),queries, keys, and values all come from the outputs of the previous decoder layer. When training sequence-to-sequence models, tokens at all the positions(time steps) of the output sequence are known. However, during prediction the output sequence is generated token by token; thus, at any decoder time step only the generated tokens can be used in the decoder self-attention. To preserve auto-regression in the decoder, its masked self-attention specifies a valid length so that any query only attends to all positions in the decoder up to the query position. Residual Connection and Layer Normalization There is also normalization layer and residual connections in between each subsequent attentions in both encoder and decoder to ensure efficiency.

E. Evaluation

The Qualitative/Quantitative analysis of generated caption can be evaluated by using NLP based Metrics such as BLEU scores. The results of these metrics shall be explained in results and observation section.

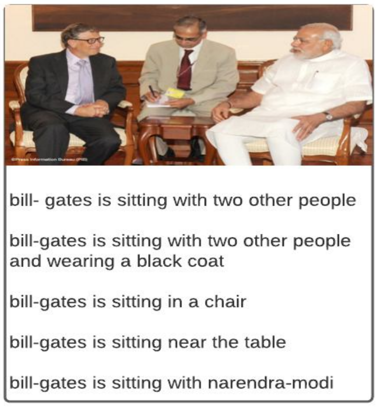
1) *Bilingual Evaluation Understudy Score:* The Bilingual Evaluation Understudy Score, or BLEU for short, is a metric for evaluating a generated sentence to a reference sentence[15]. A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0. The primary task for a BLEU score id to compare n-grams of the predicted caption with the n-grams of the reference caption and count the number of matches. These matches are position-independent. The more the matches, the better is the predicted caption.
2) *Metric for Evaluation for Translation with Explicit Ordering:* The METEOR[16] metric is designed to address some of the deficiencies inherent in the BLEU metric. The metric is based on the weighted harmonic mean of unigram precision and unigram recall. It also includes some other features not found in other metrics, such as synonymy matching, where instead of matching only on the exact word form, the metric also matches on synonyms. The metric is also includes a stemmer, which lemmatises words and matches on the lemmatised forms.

IV. EXPERIMENTAL RESULTS

A. Dataset

The dataset comprises of 362 images in total and 232 unique images of business people, popular politicians and on-screen celebrities. Most of the image has multiple captions as given in Fig. 3 describing their attributes, actions, and gender.Each image has its own unique name. The images in the dataset are obtained from various websites. Therefore it has images of various celebrities around the world with around five captions for an each image. Descriptions in the caption is mostly based on object of primary focus instead of explaining all the objects in an image.

The images were selected in accordance to their actions in the image and the captions have their dress attributes, their actions and their positions. The current study have been implemented using private dataset. However, 100 random images from public dataset called flickr 8k have been taken to increase the size of the dataset. Custom dataset was also generated for training Faster R-CNN. The custom dataset includes nameof an image and co-ordinates i.e bottom-left and top-right around the face of the celebrity and , rest two co-ordinates are calculated using basic geometric rules.



bill- gates is sitting with two other people

bill-gates is sitting with two other people and wearing a black coat

bill-gates is sitting in a chair

bill-gates is sitting near the table

bill-gates is sitting with narendra-modi

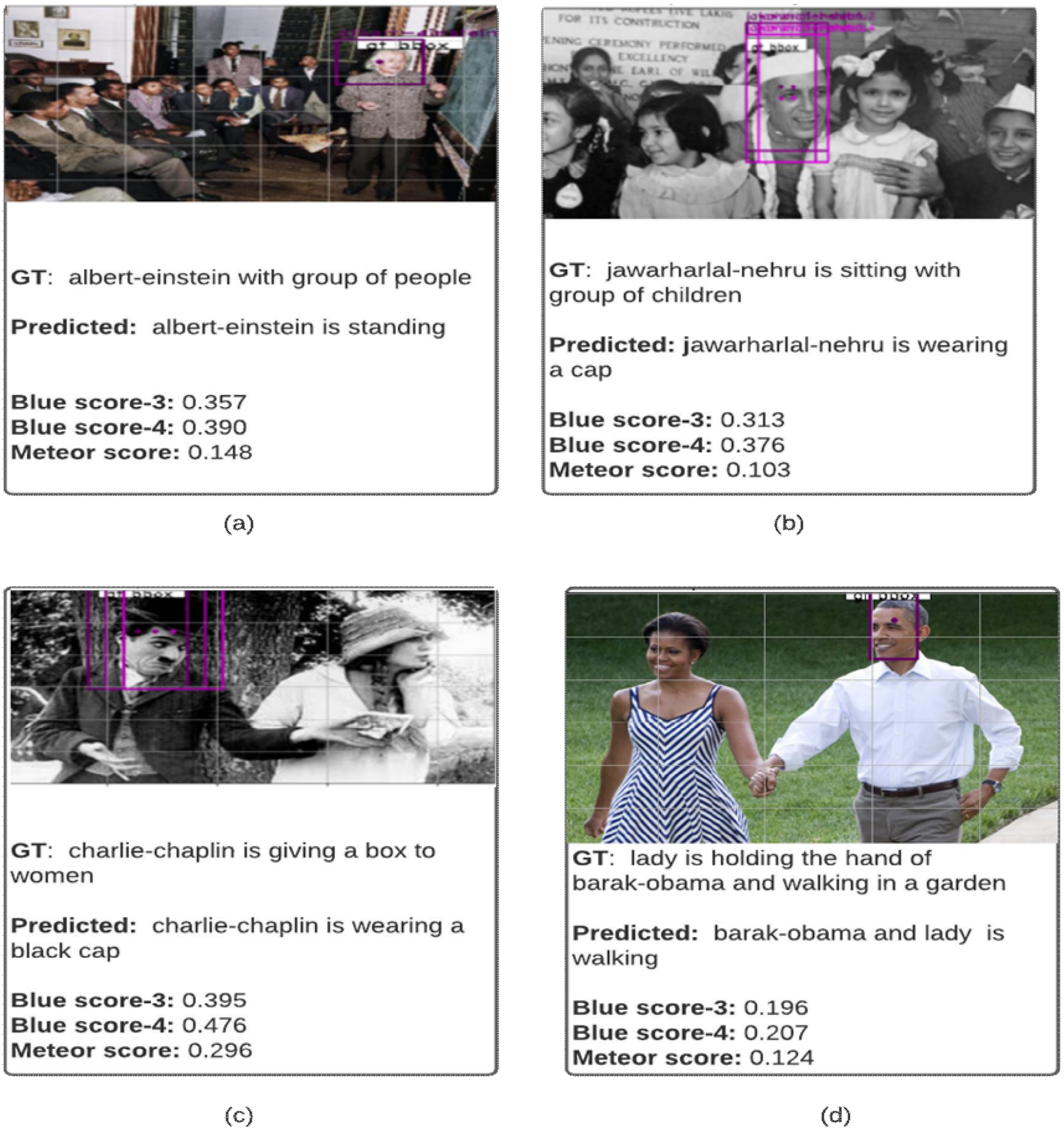
Fig. 3 sample from dataset

B. Results and Observation

The results for this work, before and after augmentation have been shown in Table.1. The output captions generated by our model are depicted in Fig. 4 and are compared against the model [1] since both of the models uses same datasets as a base. It has been observed that dataset without image augmentation is not capable to give satisfactory results. However, after augmentation of images it gives satisfactory results. Also, observed that semantics of generated captions has improved after augmenting the images and randomly selected 100 images and 500 corresponding captions from flickr dataset (to increase size of vocabulary). The direct relation between size of vocabulary and generated captions have been identified.

Model	Blue score-3	Blue score-4	Meteor score
Hemalatha et al [1]	0.33	0.23	018
Before Augmentation (our model)	0.32	0.42	0.31
After Augmentation (our model)	0.40	0.45	0.34

Table. 1 Performance of the proposed model.


gt_bbox
GT: albert-einstein with group of people
Predicted: albert-einstein is standing
Blue score-3: 0.357
Blue score-4: 0.390
Meteor score: 0.148
gt_bbox
GT: jawarharlal-nehru is sitting with group of children
Predicted: jawarharlal-nehru is wearing a cap
Blue score-3: 0.313
Blue score-4: 0.376
Meteor score: 0.103
gt_bbox
GT: charlie-chaplin is giving a box to women
Predicted: charlie-chaplin is wearing a black cap
Blue score-3: 0.395
Blue score-4: 0.476
Meteor score: 0.296
gt_bbox
GT: lady is holding the hand of barak-obama and walking in a garden
Predicted: barak-obama and lady is walking
Blue score-3: 0.196
Blue score-4: 0.207
Meteor score: 0.124

(a)

(b)

(c)

(d)

Fig. 4 Results generated by proposed model

C. Training Details

The proposed system is developed using python, Tensorflow, Keras framework and the performance of the designed system is evaluated based on Bleu score and Meteor scores.

1) No. of attention layer = 4.
2) Dimensions of Embedding Vector = 512.
3) No.heads (multi head) = 8.
4) Dropout rate = 0.1.
5) No. of epoch = 8.
6) Adam optimizer with beta$_1$=0.9 and beta$_2$=0.98.

V. CONCLUSION

In this work, image captioning model using transformer is developed along with recognizing faces. Here, Faster R-CNN as an encoder at base, and the transformer model as a decoder at base is employed. The captions generated with image and data augmentation has achieved better performance compared to caption generation without augmentation. The problem of long-range dependency have been resolved by using transformer network because it is capable to focus on particular words on either side of the word by using positional encoding and vector embedding together. Hence long-term dependency issue is resolved. The drawback of the transformer is, that it is data hungry and requires more space. Captions generated by the current model are much accurate to actual objects and activities present in an image. The current model ables to detect only one celebrity in an caption but not works better when they more than one celebrity. The model is prone to give inaccurate results while replacing appropriate noun in generated caption. The future scope may include Dense image captioning model, where multiple captions is generated locally at each segment of an image. And possibly extended to story generation from an image. We may also employ Vision Transformer[14] for feature extraction from images. Implement beam search[19] (select top k highest probabilities) instead of grid search[18] which selects only highest probability while predicting the word calculated by a softmax layer.

REFERENCES

[1] L. Abisha Anto, S. Jeevitha, M. Madhurambigai, and M. Hemalatha. ”A Semantic Driven CNN–LSTM Architecture for Personalised Image Caption Generation.” In 2019 11th International Conference on Advanced Computing (ICoAC), pp. 356- 362. IEEE, 2019.

[2] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. ”Attention is all you need.” In Advances in neural information processing systems, pp. 5998-6008. 2017.

[3] Hossain, MD Zakir, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. ”A comprehensive survey of deep learning for image captioning.” ACM Computing Surveys (CsUR) 51, no. 6 (2019): 1-36.

[4] Mishra, Santosh Kumar, Rijul Dhir, Sriparna Saha, Pushpak Bhattacharyya, and Amit Kumar Singh. ”Image captioning in Hindi language using transformer networks.” Computers Electrical Engineering 92 (2021): 107114.

[5] Lakshminarasimhan Srinivasan, Dinesh Sreekanthan, and A. L. Amutha. ”Image Captioning-A Deep Learning Approach.” International Journal of Applied Engineering Research 13, no. 9 (2018): 7239-7242.

[6] Aldabbas, Hamza, Muhammad Asad, Mohammad Hashem Ryalat, Kaleem Razzaq Malik, and Muhammad Zubair Akbar Qureshi. ”Data Augmentation to Stabilize Image Caption Generation Models in Deep Learning.” (2019).

[7] A. Kojima, T. Tamura, K. Fukunaga, Natural language description of human activities from video images based on concept hierarchy of actions, International of Computer Vision 50 (2002) 171–184.

[8] P. Hede, P. Moellic, J. Bourgeoys, M. Joint, C. Thomas, Automatic generation of natural language descriptions for images, in: Proc. Recherche Dinformation Assistee Par Ordinateur, 2004.

[9] 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: European Conference on Computer Vision,, 2010, pp. 15–29.

[10] 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) 853–899.

[11] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.

[12] J. Mao, W. Xu, Y. Yang, J. Wang, A. L. Yuille, Explain images with multimodal recurrent neural networks, arXiv preprint arXiv:1410.1090.

[13] Jiang, Huaizu, and Erik Learned-Miller. "Face detection with the faster R-CNN." In 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017), pp. 650-657. IEEE, 2017.

[14] Kolesnikov, Alexander, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer et al. "An image is worth 16x16 words: Transformers for image recognition at scale." (2021).

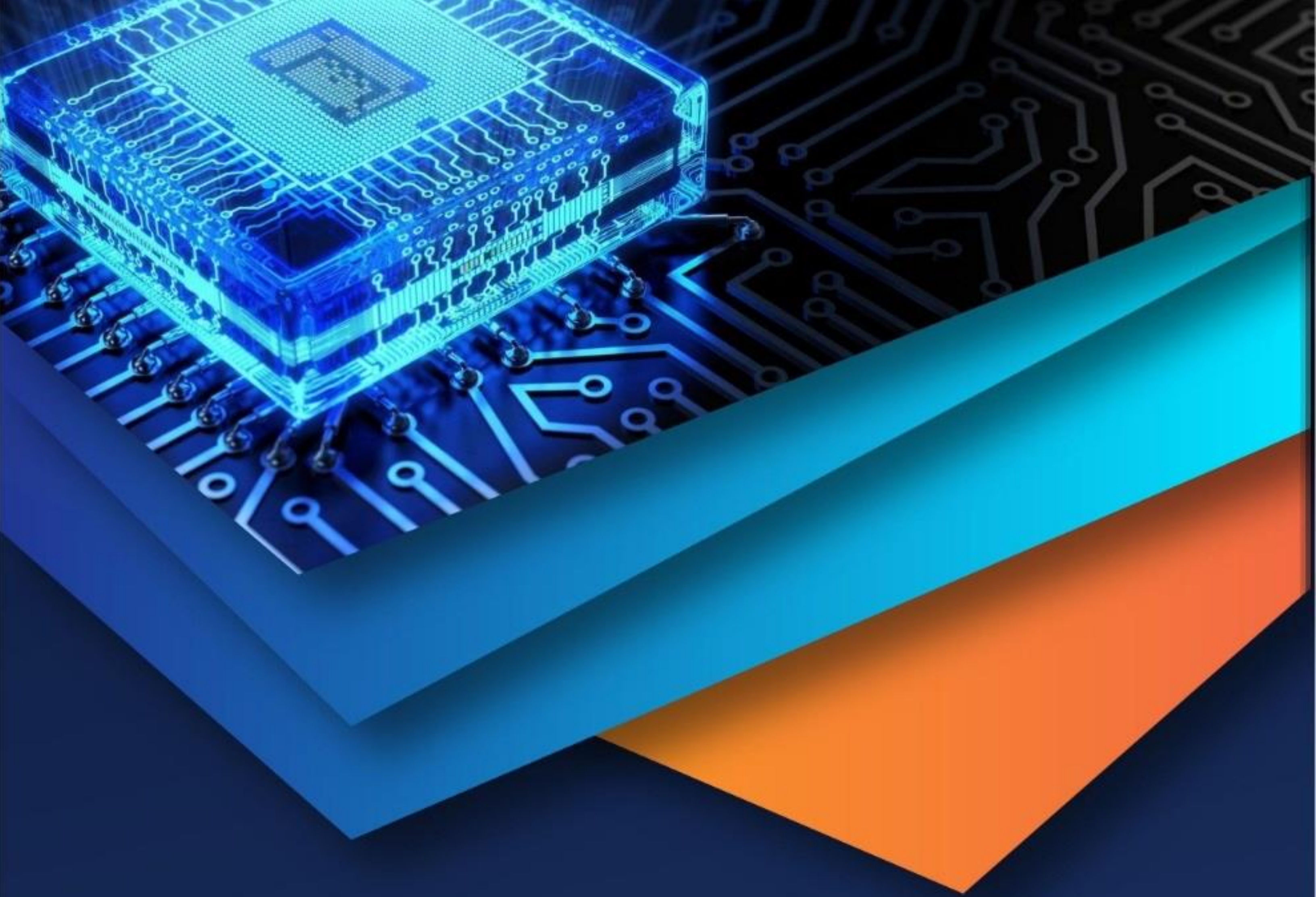
[15] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a method for automatic evaluation of machine translation." In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311-318. 2002.

[16] Lavie, Alon, and Michael J. Denkowski. "The METEOR metric for automatic evaluation of machine translation." Machine translation 23, no. 2-3 (2009): 105-115.

[17] Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the inception architecture for computer vision." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826. 2016.

[18] Liashchynskyi, Petro, and Pavlo Liashchynskyi. "Grid search, random search, genetic algorithm: a big comparison for NAS." arXiv preprint arXiv:1912.06059 (2019).

[19] Xu, Zhong-wei, Feng Liu, and Ying-xin Li. "The research on accuracy optimization of beam search algorithm." In 2006 7th International Conference on Computer-Aided Industrial Design and Conceptual Design, pp. 1-4. IEEE, 2006.



10.22214/IJRASET

INDEX COPERNICUS

45.98

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY