Fine Informative Image Captioning Using Different Deep CNN Architecture

Jansi Rani. J
Research Scholar,
Dept. of Computer Science and Engineering, Annamalai University.

Kirubagari.B
Associate Professor
Dept. of computer science and Engineering, Annamalai University.

Abstract-Every Image conveys explicit information. People can comprehend the idea of an Image on account of communication with the world yet a machine can't comprehend the idea of an Image all alone. Image Captioning is a fundamental activity that includes semantic comprehension of images and the capacity to make clarification sentences with exact and right words. To create the Image Captioning we need to blend Computer Vision and Natural Language Processing (NLP). The main goal of this task is attaining fine informative captioning of an image, where better visual representationis extracted and mapping with better text. Diverse Deep Convolution Neural Network (DCNN) Architectures like VGG Net 16 and Inception V3 are utilized with flickr8K Dataset to achieve better visual representation. Long Short-Term Memory (LSTM) is utilized to produce fine informative captioning from visual representation. We train and developfeature extracted images utilizing Inject Merged architecture model tuned with Adam optimizer and categorical-cross entropy loss is calculated. The acquired outcome is evaluated by benchmark merits BLEU, GLUE, Meteor and human-created sentences. The output of diverse DCNN models is compared and discussed a best captioning as Fine Informative Captioning.

Keywords: Deep Learning; Image Captioning; CNN; LSTM; VGG Net16; InceptionV3.

I.INTRODUCTION

In development of artificial intelligence, Image captioning is a challenging task. the several authors have explored in this field to accomplish fine captioning that is utilized for a few applications, for example, image finding, self-driving vehicle, and menial helpers, suggestions in editing application, image retrieval, biomedicine, business, military, training, outwardly crippled individuals, web-based media and many other natural language processing. Scientists are begun working on visual perception in images,it became clear that solely providing the names of the objects recognized doesn't build such a decent impression as a full human-like description. There are a few variations and combination of different techniques are utilized in deep neural networks in image captioning. Numerous works were at that point attempted to get a lot of correct descriptions and construct machines expect like people. However, machines still lack the natural means of human communication and this continues to be a difficult task in image captioning.

Image captioning is that the route toward generating textual descriptions for an image,found on objects and action. Image caption generator requires the two strategies of computer vision to understand the content of the image and a language model from the field of natural language processing to turn the image into informative words. Dataset are fine-tuned with Preprocessed Techniques and these features are isolated with two deep learning trained architecture of VGG 16 and the Inception V3 model by transfer learning method. Extracted features are utilized for Image Input and text of each image are preprocessed and utilized for Text Input. LSTM is utilized to predict the captioningin a sentence structure. The LSTM is a unique sort of RNN that is fit for learning long-term dependencies.We trained and tested on our proposed “Inject Merged model “with two kinds of feature extraction and results are compared.

The paper is coordinated as follows: Section II clarifies the connected work of Image Captioning, Section III covers pre-processing of datasets. SectionIV portrays the proposed model. Segment V depicts results and examinations with Evaluation Metrics. At long last closes the paper.

I. Related Works of Image Captioning

In the previous year 2005 to 2010, Major works were accomplishedwith Computer Vision and afterward began mapping the identified objects with words to shape significant description. Various kinds of captioning were sequenced. In Template-based Image Captioning, Farhadi et al [1] utilized a trio scene of components like object, activities and attributes to fill the format spaces for creating Image Captions. Li et al [2] extricate the expressions related toobjects, attributes and connections. Kulkarni et al [3] gather the articles, ascribes with relational words before fills the template gaps. Template-based techniques are predefined can't produce variable-length captions however create syntactically right captions [4, 5, 6, 7]. In retrieval-based procedures, captions are regained from a collection of existing captions.Captions can be recuperated from visual space and multimodal space [8]. Visual space methods first find ostensibly relative images with their captions from the trainingdataset and subsequently depict the engravings, which produce general and linguistically right captions yet can't make semantically right captions. Multimodal space strategies join both image and text for making Image caption.

Novel subtitle strategies utilize both visual space and multimodal space [9]. This technique doesn't depend on any additional templates, designs, or imperatives. It relies upon the undeniable level image features and word representations gained from deep neural Networks. The language models have restrictions to deal with a lot of information and are ineffective to work with long-term memory [10].

The Image captioningmethodology of encoder-decoder procedure functions as a simple end-to-end-manner. It requires an image model and a language model. Image features are extricated from hiddenactivations of CNN. Scene types were predicted. The image objects and their relationship were identified by CNN for Image representations. BatchNormalization is utilized in the output of the last layer of CNN [11]. Batch Normalization is a technique for training very deep neural networks that normalizes the inputs to a layer for each mini-batch. This stabilizes the learning process and drastically diminishing the amount of training epochs needed to traindeepnetworks.Recurrent Neural Network (RNN) is utilized to decode this image representation into natural languagedepictions [12]. LSTM is utilized to have a track of the Image Objectthat already have been stored in the depiction. A LSTM layer comprises of recurrently connected blocks referred to as memory blocks. These memory blocks are regularly considered as a differentiable version of the memory chips in a digital computer. Every memory blocks contains one or more repetitively connected memory cells and three multiplicative units are the input ,output units and forgets gates that gives persistent analogs of write, read and reset activities for the cells [13]. The underlying condition of a LSTM involved the image data, the following words are created dependent on the present status and past hidden action, this cycle proceeds until it gets the end badge of the sentence. Producing long length sentences was challengeable for LSTM [14, 15] guided LSTM (gLSTM) is utilized to create long sentence. The Bidirectional LSTM based strategy is equipped for producing relevantly and semantically rich image captions, by two separate LSTM Networks, which convey the data to and from in the neural network [16]. It also utilizes both past and future context information to learn long term visual language interactions.

III. METHODOLOGY AND MATERIALS

3.1 Pre-processing of Dataset

Information could be in such various structures: Structured Tables, Images, Audio records, Videos and so forth. Some enormous datasets contain large number of rows and columns. Machines don't comprehend the free text, image or video information for what it's worth, they get 1s and 0s. Along these lines, we need to preprocess the dataset. Deep learning models can't get trained just by raw information. We utilized Flickr 8K dataset for the process of image caption generation. It is a labeled dataset comprising of 8000 photographs with 5 captions for every photograph [17]. The images are obtained from the Flickr website.Dataset pre-processing allows resizing and normalizing the images

from the Flickr8k dataset and creating the image encoding from the resized images using state-of-the-art Convolutional Neural Networks. Likewise, the Descriptions are fine-tuned by making the word-to-id references of the relative multitude of words that encode the image caption. From flickr8k Dataset, 6000 photos are utilized for training, 1000 photos are utilized for creating and 1000 photos are tested.

3.2 Our Proposed Model

Our proposed model depends on Encoder-Decoder Architecture fills as end to end manner accordingly, it peruses an image and extracts global image features and decodes the features into a sequence of words (Fig.1). In the encoder part, CNN can create a rich Image features from the raw input by implementing it to a fixed-length vector representation [18]. LSTM is utilized to create significant text in the Decoder Part.

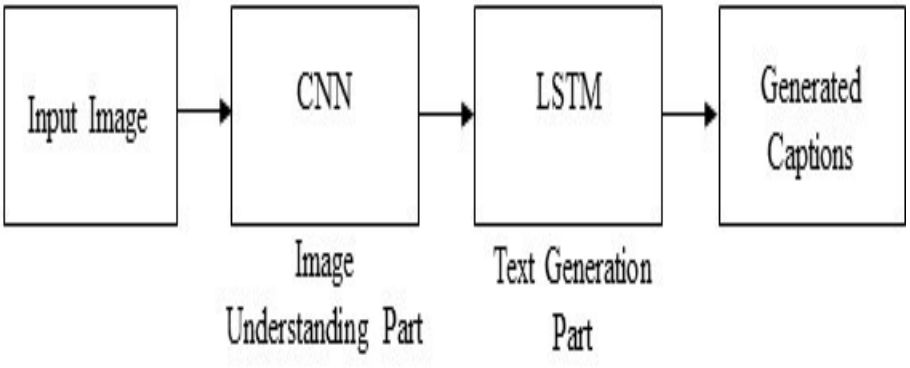

Fig. 1: Encoder-Decoder Architecture

Our Inject and Merge Model is a simple encoder and decoder Architecture is a deep learning model where extracted image features and text features are encoded into a fixed vector representation. In the inject model (Fig. 2), the encoder first encodes the image into a fixed-length vector representation. The model combines the encoded form of the image with each word from the text description generated.

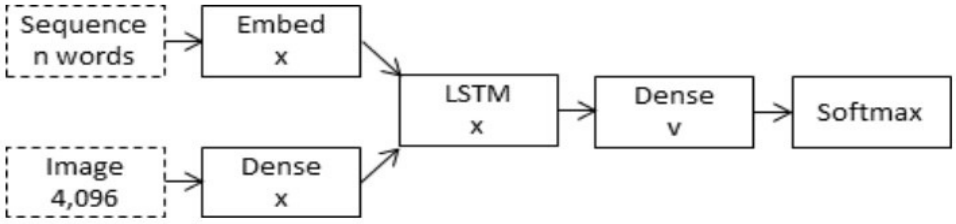

Fig. 2: Inject Architecture

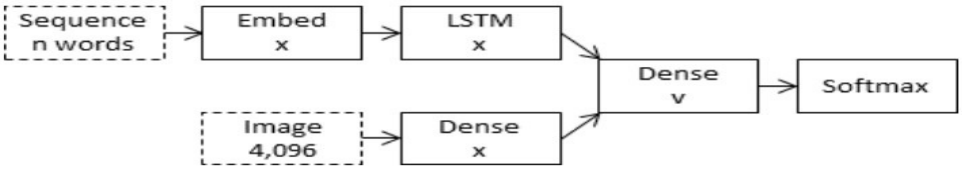

Fig. 3: Merge Architecture

The decoder acts as a text generation model that uses a sequence of both image and word information as input to generate the next word in the sequence [19]. For example; an image I have taken as Input and is trained to maximize the likelihood p(SlI) of producing a target sequence of words S={S1, S2….} where each word S1 comes from a givendictionary, that describes the image adequately [20].

In merge model (Fig. 3) it consolidates both the encoded type of the image contribution with the encoded type of the text-based depiction. The mix of these two encoded inputs is then utilized by a manageable decoder model to produce the following word in the succession [21].

We train the Inject and Merge engineering-based model with a CNN as the Encoder of input image and a LSTM as the encoder of text sentence. The model is executed utilizing the Python SciPy environment with Kera's 2.4.3. Tensor flow library is introduced as a backend for the Keras system for making and preparing deep neural networks. Tensor Flow is a deep learning library created by Google. It provides a heterogeneous platform for the execution of algorithms even it can be run on low-power devices like mobile as well as large-scale distributed systems containing thousands of GPUs. Our deep learning model is prepared on the Nvidia Geoforce GTX 1660 illustrations handling unit which has 640 Cuda cores.

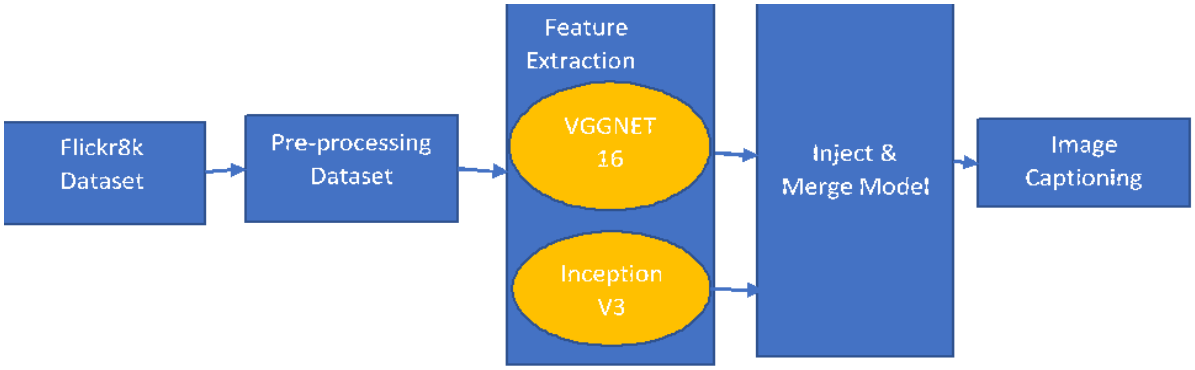

Fig. 4: Image Captioning Model

The photo features are pre-computed utilizing the pre-trained model and saved. These features are then loaded in our model as the interpretation of a given photo to reduce the redundancy of running each photo through the network every time [22]. The typical learning architecture of the model is appeared in Fig. 4.LSTM portion of the model is set up to portray each word in the sentence after it has seen image similarly as each and every previous word. For each image caption, we utilized the two images additionally to demonstrate the start and completing of the gathering. Whenever a stop word is capable it stops making a sentence and stamps it as the completion of the text. The Loss work for a model is resolved using, where I tend to the information of input image and S tends to the generated caption. N is the length of created sentence. pt and St address probability and expected word at the time t exclusively (eq .1). During the path toward getting training, we have endeavored to restrict the loss function.

$$L(I, S) = -\sum_{t=1}^{N} \log p_t(St) \quad (1)$$

We train and tested the model with the following variations and compared the results:

1. VGG16 model as the image encoder and an LSTM as the sentence encoder.
2. Inception V3 model as the image encoder and an LSTM as the sentence encoder.

3.3 DCNN architecture

3.3.1 VGG 16 Model

The VGG is a convolutional neural organization that comprise of 16 layers which have samples of 2 convolutional blocks, one dropout layer and a FC(fully connected layer) toward the end. The dropout layer present to diminish overfitting of the training images, as the model setups catch on quickly. These are prepared by a Dense layer to

create 4096 vector representation of the information and provided to the LSTM layer without the SoftMax layer [23].

The contribution to the organization is an image of measurements (224, 224, 3). The initial two layers have 64 channels of 3*3 channel size and a similar padding. At that point there are 2 convolution blocks of channel size (3, 3) and 256 channels. From that point forward, there are 2 sorts of 3 convolutions blocks and a maximum pool layer. Each has 512 channels of (3, 3) size with a similar padding. This input is then passed to two convolution layers. In these convolution and max-pooling layers, the channels we use are of the size 3*3. In a portion of the layers, it likewise utilizes 1*1 pixel which is utilized to control the quantity of information channels. There is a padding of 1-pixel done after every convolution layer to forestall the spatial element of the image appeared in Fig.5.

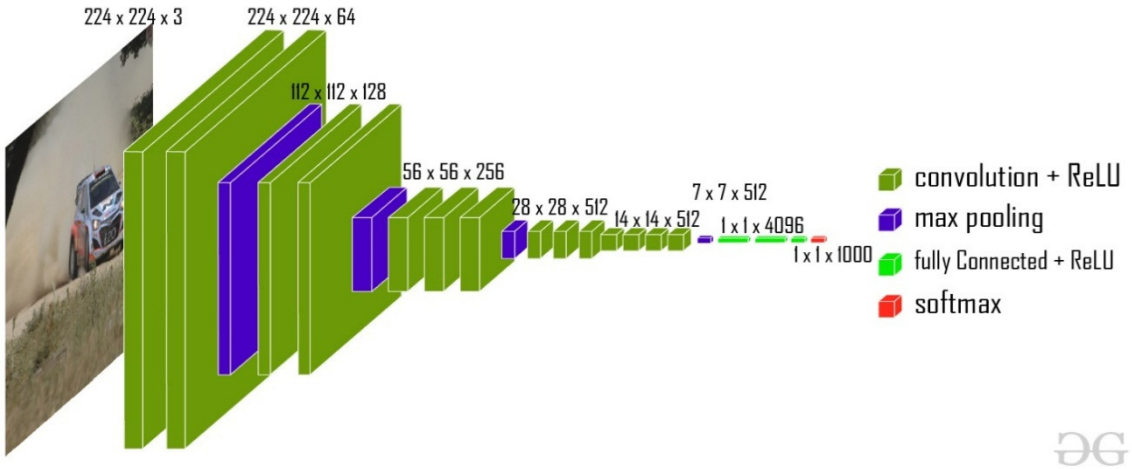

Fig. 5: VGG 16 Architecture

3.3.2 Inception V3 model

Inception v3: Inception-v3 (is a convolutional neural organization that has been trained on an enormous number of images in the ImageNet information base). The neural network is 48 layers profound and can order images into 1000 article classes. Inception v3 has a data input size of 299-by-299. It comprises of convolution layer, the Inception family that makes a few enhancements including utilizing Label Smoothing, factorized 7 x 7 convolutions, and the utilization of a helper classifier to proliferate label data lower down the neural network system (alongside the utilization of group standardization for layers in the side head) [24].

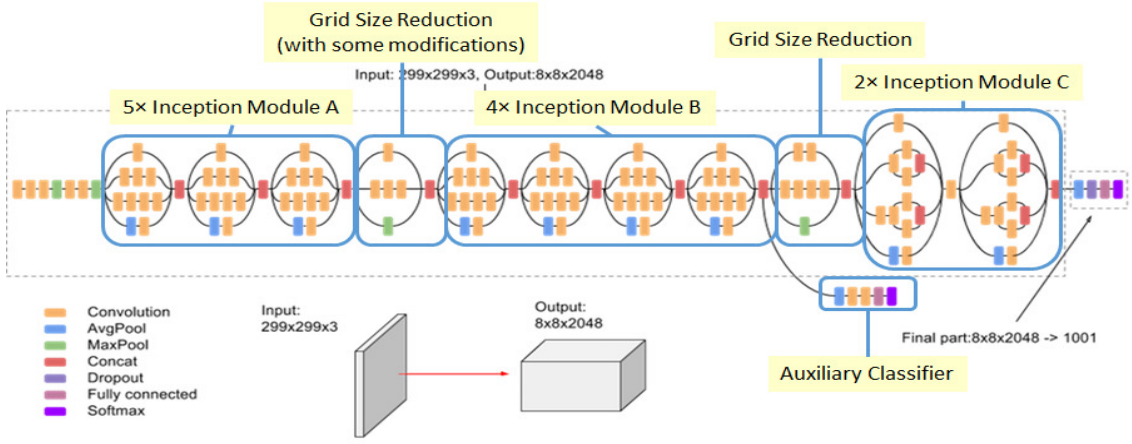

Fig. 6: Inception V3 Architecture

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Captioning of the Image is assessed by BLEU, GLEU, METEOR and Human-Generated sentences. BLEU (Bilingual Evaluation Understudy Score) is the most mainstream technique for estimating the comparability among sentences and broadly used to survey the nature of the produced caption. The processed scores are arrived at the midpoint of BLEU score is the mathematical mean of n-gram exactness scores increased by Brevity Penalty (BP) for more limited sentences:However, syntactical accuracy isn't considered here. The presentation of the BLEU metric is changed relying upon the quantity of reference interpretations and the size of the created text. BLEU is well known on the grounds that it is a pioneer in the programmed assessment of machine-translated content and has a sensible connection with human decisions of value [25]. However, it has a few constraints; for example, BLEU scores are acceptable just if the produced text is short. There are a few situations where an expansion in BLEU score doesn't imply that the nature of the produced text is acceptable.

The enhanced benchmark metric of General Language Understanding Evaluation (GLUE) over the BLEU score proposed by Google in their 2016 paper" Googles Neural Machine Translation System Bridging the Gap among Human and Machine Translation". This measurement is utilized to assess the nature of the machine-translation frameworks, however it very well may be utilized for better assessing the nature of image caption than BLEU. METEOR (Metric for Evaluation of Translation with Explicit Ordering) [26] is a measurement for the assessment of machine-interpretation output. The measurement depends on the harmonic s mean of unigram exactness and recall, with recall weighted higher than accuracy. It additionally has a few highlights that are not found in different measurements, for example, stemming and synonymy coordinating, alongside the standard accurate word coordinating. The measurement was intended to fix a portion of the issues found in the more famous BLEU metric, and furthermore produces a decent connection with human judgment at the sentence or section level. This contrasts from the BLEU metric in that BLEU looks for relationship at the corpus level METEOR, similarly as GLEU, is intended for the assessment of the nature of the machine translation framework. The nature of image caption relies upon the appraisal of two primary viewpoints: adequacy and fluency. An assessment metric necessity to zero in on a different arrangement of phonetic highlights to accomplish these angles. Be that as it may, normally utilized assessment measurements looking only particular extraction (e.g., lexical or semantic) of languages. Our Generated captions additionally contrasted and the most extreme Level of human expected caption [27]. The exactness is determined by the score 0 to 1, where the most exceedingly awful image caption is underneath 0.5 and the best image caption is above 0.5.

An undeniable application is created which accepts a image as an input data and output as caption related with it. This is constructed utilizing a Deep CNN learning model comprises of an encoder-decoder pipeline. A CNN encoder model is utilized with pre-trained loads. A pre trained Bidirectional LSTM is utilized as a decoder. Adam optimizer is utilized to tune the inject merged model, whereas Adam optimization algorithm is an extension to stochastic gradient descent that has recently seen broader adoption for deep learning applications in computer vision and natural language processing.

While training the Deep learning Inject and Merge model, we understood that in the function of picking the best model for the assignment, a few models are should have been trained and analyzed against regular measurement such BLUE, GLUE, METEOR and Human-Generated Sentences. The evaluation on measurements like Training time, training loss, validation accuracy, validation loss, training accuracy of the best model is likewise gathered Fig. (7,8). We empowered early halting to keep the model from over-fitting the dataset and limit training time. The performance Metrics of VGGNET 16 and Inception V3 appeared in (Table.1) and the reference chart are appeared in (Fig. 13). The benchmark Caption of the image and our best caption are presented in Fig 9-12.

Table.1: Evaluation Metrics Scores

Metrics	VVG NET16	Inception V3
BLEU	0.6345	0.6779
GLEU	0.6526	0.6781
METEOR	0.6534	0.6854
HUMAN GENERATED SENTENCE	0.8000	0.8125

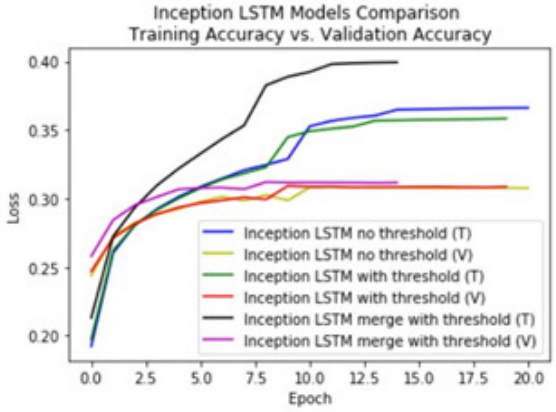

Fig. 7: Inception V3Comparison of Training Accuracy vs Validation Accuracy.

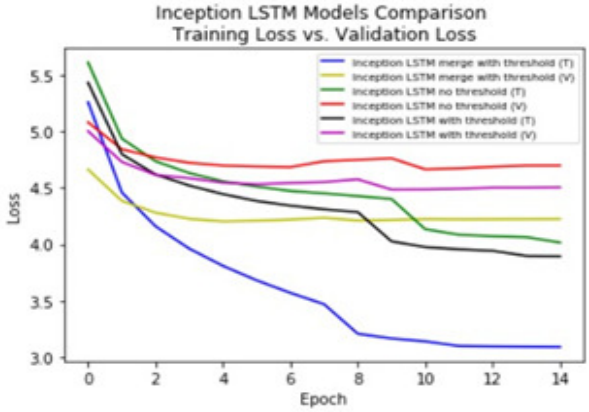

Fig.8. Inception V3 Comparison Training Loss vs Validation Loss

Fig. 9

Baseline Caption: several people are standing Near fish pond.

Our Best Caption: There are several people and children looking into water with blue tiled floor and goldfish.

Fig. 10

Baseline Caption: boy in yellow is riding scooter on The street

Our Best Caption: A young boy wearing an orange helmet Scatting on the road.

Fig.11Fig.12

Baseline Caption: climber climbing an ice wall.

Our Best Caption:A man in blue jacket and black pants climbing on frozen ice wall with two picks.

Baseline Caption: black and white bird Standing on hand.

Our Best Caption: A small bird standing on someone hand with seeds.

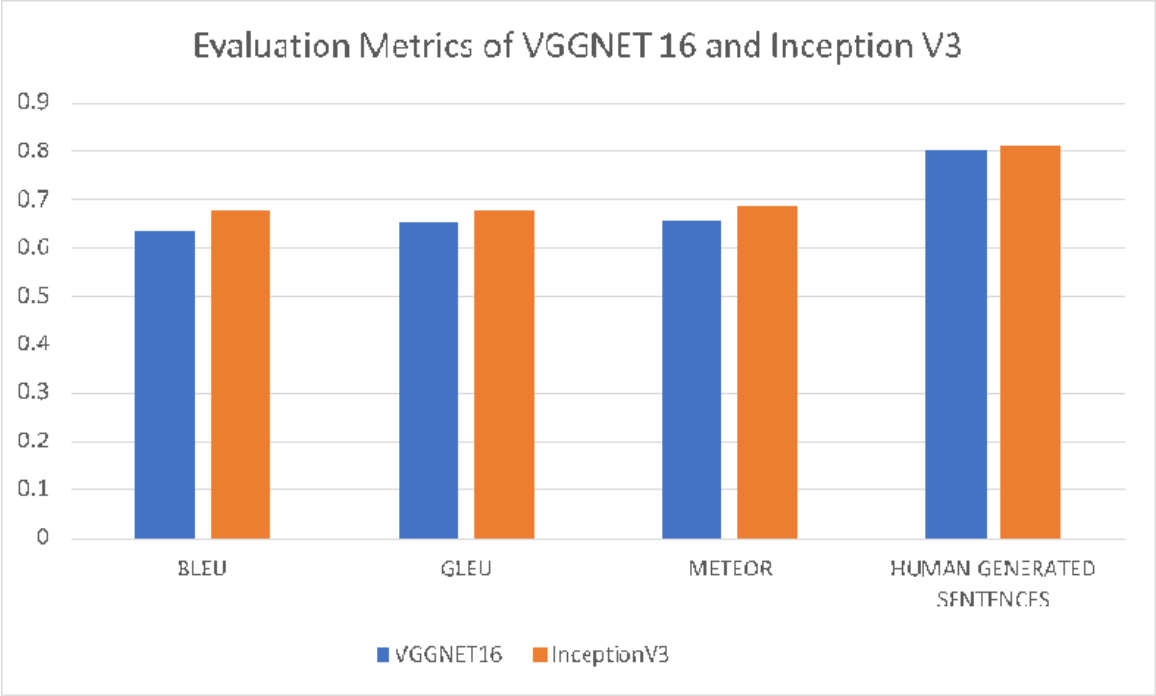

Fig 13

V. CONCLUSION

In this paper, we carried out a Deep learning approach for the caption of pictures. The successive API of Kera's was utilized with Tensor Flow as a backend to execute the DCNN architecture of VGG 16 and Inception V3 to accomplish a viable BLEU, GLEU, METEOR and human created Sentence score. An ideal match brings about a score towards 1, while an ideal crisscross outcome in a score towards 0. The Output of Inception V3 (Table.1) ready to get fine informative captioning, contrasted with VGG NET 16. The execution of System Configuration used to train the model rapid computation and Low accuracy loss. Later on, needs to be an improve accomplish better execution by utilizing word vectors on a lot bigger corpus of information, such as news stories and other online wellsprings of information.

References

[1] A. Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In European conference on computer vision. Springer, 15–29.

[2] Siming Li, GirishKulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 220–228.

[3] GirishKulkarni, VisruthPremraj, SagnikDhar, Siming Li, YejinChoi, Alexander C Berg, and Tamara L Berg.2011. Babytalk:Understandingandgeneratingimagedescriptions.InProceedingsofthe24thCVPR.Citeseer.

[4] Ahmet Aker and Robert Gaizauskas. 2010. Generating image descriptions using dependency relational patterns. Proceedings of the48 th annual meeting of the association for computation linguistics .Association for Computational Linguistics,1250–1258.

[5] WilliamFedus,IanGoodfellow,andAndrewMDai.2018.Maskgan:Bettertextgenerationviafillinginthe_.*arXiv preprintarXiv:1801.07736*.

[6] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In ICML, 2015.

[7] PolinaKuznetsova, Vicente Ordonez, TamaraLBerg,andYejinChoi.2014.TREETALK:Compositionand Compression of Trees for Image Descriptions. *TACL* 2, 10 (2014),351–362.

[8] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving image- sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision*. Springer, 529–545.

[9] RyanKiros, RuslanSalakhutdinov, andRichZemel.2014.Multimodal neurallanguagemodels.In*Proceedingsofthe 31st International Conference on Machine Learning (ICML-14)*.595–603.

[10] RafalJozefowicz,OriolVinyals,MikeSchuster, NoamShazeer,andYonghuiWu.2016.Exploringthelimitsoflanguage modeling. *arXiv preprint arXiv:1602.02410*(2016).

[11] OriolVinyals, Alexander Toshev, SamyBengio, and DumitruErhan. 2015. Show and tell: A neural image caption generator.In*ProceedingsoftheIEEEconferenceoncomputervisionandpatternrecognition*.3156–3164.

[12] TomášMikolov,MartinKarafiát, LukášBurget, JanČernocky`, andSanjeevKhudanpur.2010.Recurrentneuralnetworkbasedlanguagemodel.In*EleventhAnnualConferenceofthe InternationalSpeechCommunicationAssociation*.

[13] Graves. A, Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM networks. In Proceedings. 2005 IEEE International Joint Conference on Neural Networks.

[14] Dim P Papadopoulos, Alasdair DF Clarke, Frank Keller, and Vittorio Ferrari. 2014. Training object class detectors fromeyetrackingdata.In*EuropeanConferenceonComputerVision*.Springer,361–376.Kilickaya, Mert. Re-Evaluating Automatic Metrics for Image Captioning. Https://Www.aclweb.org/, Association for Computational Linguistics, Apr. 2017, https://www.aclweb.org/anthology/E17-1019/

[15] XuJia, Efstratios, Gavves,BasuraFernando,andTinneTuytelaars.2015.Guidingthelong-shorttermmemorymodel forimagecaptiongeneration.In*ProceedingsoftheIEEEInternationalConferenceonComputerVision*.2407–2415.

[16] Cheng Wang, Haojin Yang, Christian Bartz, and ChristophMeinel. 2016. Image captioning with deep bidirectional LSTMs.In*Proceedingsofthe2016ACMonMultimediaConference*.ACM,988–997.Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., &Wojna, Z. (2016).

[17] MicahHodosh,PeterYoung,andJuliaHockenmaier.2013.Framingimagedescriptionasarankingtask:Data,models and evaluation metrics. *Journal of Artificial Intelligence Research* 47 (2013),853–899.

[18] Colah(2015),Understanding LSTM Networks. Retrieved from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[19] IlyaSutskever,OriolVinyals, andQuocVLe.2014.Sequence tosequencelearningwithneuralnetworks.In*Advances in neural information processing systems*.3104–3112.
https://Www.aclweb.org/, Association for Computational Linguistics, 2018, https://www.aclweb.org/anthology/P18-1238/

[20] Fang, H. (2014, November 18). From Captions to Visual Concepts and Back. Retrieved from https://arxiv.org/abs/1411.4952v3.
MicahHodosh,PeterYoung,andJuliaHockenmaier.2013.Framingimagedescriptionasarankingtask:Data,models and evaluation metrics. *Journal of Artificial Intelligence Research* 47 (2013), 853–899.

[21] Zhou Ren, Xiaoyu Wang, Ning Zhang, XutaoLv, and Li-Jia Li. 2017. Deep Reinforcement Learning-based Image CaptioningwithEmbeddingReward.In*ProceedingsoftheIEEEconferenceoncomputervisionandpatternrecognition (CVPR)*.1151–1159.

[22] Kishore Papineni, SalimRoukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machinetranslation.In*Proceedingsofthe40thannualmeetingonassociationforcomputationallinguistics*.Association for Computational Linguistics,311–318.

[23] Karen Simonyan, Andrew Zisserman 2015. Very Deep Convolutional Networks for Large Scale Image Recognition. In ICLR Conference.

[24] *Milton-Barker, Adam. 2019* "Inception V3 Deep Convolutional Architecture For Classifying Acute Myeloid/Lymphoblastic Leukemia". intel.com. Intel.

[25] Kishore Papineni, SalimRoukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluationof machinetranslation.InProceedingsofthe40thannualmeetingonassociationforcomputationallinguistics.Association for Computational Linguistics,311–318.

[26] AbhayaAgarwalandAlonLavie.2008.Meteor,m-bleuandm-ter:Evaluationmetricsforhigh-correlationwith humanrankingsomachinetranslationoutput.In*ProceedingsoftheThirdWorkshoponStatisticalMachineTranslation*. Association for Computational Linguistics,115–118