2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) | 978-1-6654-1272-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICCS51141.2021.9432245

Generating Caption for Image usingBeam Search and Analyzation with Unsupervised Image Captioning Algorithm

Prashant Giridhar Shambharkar
Department of Computer Engineering
Delhi Technological University
New Delhi, India
prashant.shambharkar@dtu.ac.in

Priyanka Kumari
Department of Computer Engineering
Delhi Technological University
New Delhi, India
priyankakumari_2k17co244@dtu.ac.in

Pratik Yadav
Department of Computer Engineering
Delhi Technological University
New Delhi, India
pratikyadav_2k17co236@dtu.ac.in

Rajat Kumar
Department of Computer Engineering
Delhi Technological University
New Delhi, India
rajatkumar_2k17co255@dtu.ac.in

***Abstract*—In today's world of social media, almost everyone is a part of social platform and actively interacting with each other through internet. People on social media upload many pictures on their social media accounts with different captions. Thinking about the appropriate caption is a tedious process. Caption is important to effectively describe the content and meaning of a picture. Caption describes the image in meaningful sentences. A model for image caption generator can be built which is used to generate caption for images of different types and resolutions. Image captioning model which is used to generate caption in language that is understandable by a human being for the input images. CNN(convolution neural network) and RNN(recurrent neural network) is used using the concept of encoder-decoder to build this model. As CNN is used for image feature extraction purpose where only the important features or the important pixels, if the image is considered in the form of matrix of pixels, which are extracted from the resultant image, instead of CNN model, other pre-trained imagenet models which have higher accuracy will be used and their results are then compared by using BLEU score metric for comparison. For the prediction of captions, beamsearch method and argmax method is used and compared. The above discussed supervised image caption model is also comparedwith the built unsupervised image captioning model.**

The flickr8k dataset and then MSCOCO dataset are used to train and test the model. This model if implemented with the mobile application, which can be very useful for differently abled people, who completely rely on the assistance of text-to-speech feature.

***Index Terms*—Image caption generator, deep learning, CNN, RNN, LSTM, supervised and unsupervised learning, beam search, Argmax.**

I. INTRODUCTION

Image caption generator being a very famous and accepted field of research and study defines a way to build a model which can generate an English language based sentence or caption that is understandable by a human brain i.e., the sentence will be both syntactically and semantically correct. The resultant model will be able to describe the image precisely to a human brain.

In recent times, the studies and research related to captioning an image or providing the image with a meaningful sentence describing the image using some advanced deep learning techniques and algorithms and the concept of computer vision have become immensely popular and advanced.

Getting the machine ready to automatically describe the objects in the picture next to it or given on the display can be a difficult work to do, but its effect is immense in many field. As an example it can be used for the visually impaired people and can be helpful to them and can act as a guidance system for their day to day life. It does not only describe the image but also make user understand about it in an understandable language.

It is the very reason, the animation field is taken seriously in this generation. This paper has the scope of copying or mimicking the brain's capability to describe the image using natural meaningful sentence and analyze various process and due to this it is a very good problem for the field of artificial intelligence which included heavy deep learning including concept of computer vision and it has resulted in many organizations studying the concept of image caption generator.

Image captioning is often used for a wide range of cases used as a way to help blind people use text to speech with real-time output about the merging of the camera feed and the particular environment, to improve public awareness by converting photo captions into social feeds as speech mes-

Identify applicable funding agency here. If none, delete this.

sages. Captions for each image on the web can create quick and precise image search and targeting. For robots, the agent's natural view is often given context by the display of natural language in the captions of images in the camera's center space.

This paper has used the encoder-decoder concept and methodology using CNN-RNN pair of architecture with beam search [1]. Here, RNN has used to provide input as the resultant optimized output of the CNN model instead of LSTM (Long Short Term Memory) to avoid the vanishing gradient problem. LSTM has the problem of vanishing gradient if the value of the parameter weight is taken to be less than 1, due to which the training of the model is improper and the resultant output is not as accurate as required. So, RNN model is used instead of LSTM where the value of weight is taken to be exactly equal to 1, if its value is greater than 1, the problem of overfitting may arise.

In this paper instead of building the CNN structure from the scratch for the optimization of the input image which is present as the matrix of pixels, various pre- trained models like Inceptionv3 and VGG16 are trained on heavy dataset resulting in a higher accuracy or lower validation loss/loss and compared their results for accuracy using BLEU (Bilingual Evaluation Understudy) scores. BLEU score is the metrics or standard to measure and identify how much understandable a sentence is to the human brain.

After covering the supervised methodology for image caption generator, we have compared the unsupervised manner for image caption generator for the accuracy of the results. The summary of our contribution to this paper:

- Built model for image caption generator with beam search using pre-trained model like VGG16 and Inceptionv3 that are already trained on huge datasets for higher accuracy.
- To resolve the "out of memory" issue, the size of batch during the training of the dataset is reduced.
- As the nature of the algorithm for image caption generator is stochastic, this may lead to slightly different results every time we run the model for results. So, random seed is set to generate same results everytime.
- We have used higher value of k in beam search for more efficient and accurate results with less validation loss/loss.
- In this paper, we compared the results of supervised and unsupervised methodology of image caption generator

II. RELATED WORK

This particular section includes investigation of the part of the work that has recently been tried in this difficult area. Pre-image captioning techniques depend on the substitutionof a productive production model for text production in a common language. Farhadi et al. (2010) [2], use trios next to a predetermined structure to produce text. They discriminate against the multi-name Markov Random Field to predict the measurement of trios. Kulkarni et al. (2011) [8], object is separated from the image, anticipate a multitude of features and related words (local data contradicting different articles) for each object, create a diagram called Basic Random Field and construct descriptive captions using marks and order. Furthermore, in the study, language parsing based model which are more efficient and powerful is also been studied and used [4, 5, 3, 6, 7]. These methods are inaccessible as they neglect to show a hidden subtitle regardless of whether each item was available in the configuration details. Likewise, the issue of format is their formal assessment.

For approaching and resolving this problem, the model is being implemented and used with the deep neural network, which has the vector space shared by both the image and the caption. Here, the Convolution Neural Network has been used as a typical approach[11] along with recurrent neural network approach [10] to the problem and it develop a number of highlights that are used in sequence modeling process that studies the model for natural language to produce an understandable language showcase. We are different from their killings in order to improve the stable conditions. The Show, Attend and Tell uses new improvements in machine description and article recognition to introduce a thought-based model that looks at a few ”spots” in a picture while making transcripts. They removed the highlights from the lower layer of convolution in contrast to the earlier layers, bringing the vector element 14 × 14 × 512 in each image. We defined the number as the ”196” spots in the picture, each with a vector of 512 objects. These 196 locations are included in the drafting process. Using visual considerations, the model had the option of finding out how to adjust its visibility to the key elements in the images when making subtitles. They introduced two instruments of thought, a "critical" object for processing prepared with back-to-back techniques; and a "solid" stochastic research program designed to maximize the hypothetical variability. Adding consideration improves the type of documentation produced, but at a higher cost of additional teachable loads. In addition, the preparation and exposure of the installed takes a ton of computer time, appropriately bringing this useless approach to the ever-present programs in the shopper's wireless devices. Karpathy and Li (2015) [12] introduced a method that uses accessible image data sets for displaying their sentences to gain expertise through multi-purpose connections among display information and descriptive language.

The planned model is defined by integrating Convolution Neural Networks into the display information areas, bidirectional Recurrent Neural Networks in the language, and the strategic goal of adapting these approaches. The design of the Multimodal Recurrent Neural Network, in these developed lines, uses expert arrangements to draw the required graphic text. The orderly model produces the best in the class bringing the research found into Flickr8K, Flickr30K and MSCOCO [13] datasets. As there are various improvements and enhancements in the field of the frameworks, Etienne et al (2016) [14] introduced another approach to improvement, small image go through independent sequence training, which is a modified form of reinforce [15]. Thus, the problem is resolved and help to develop their own model of inseparable test standards such as BLEU, ROUGE, etc., rather than being a coincidence.

III. METHODOLOGY AND ARCHITECTURAL ANALYSIS

A. Datasets and Evaluation

This part of the paper helps in understanding the image captioning and open source database and the various method for it. The important parts or element for all the inventions that has be created by humans are feed up by the data, statistics and obviously the system power. These keys are important foreach other. There is a fact that the data can make a model more powerful, reliable and more effective for its usage. The showing of the image on the display is somewhat familiar with description of the machine, and its testing strategy extends to the machine description to create its own unique test rules.

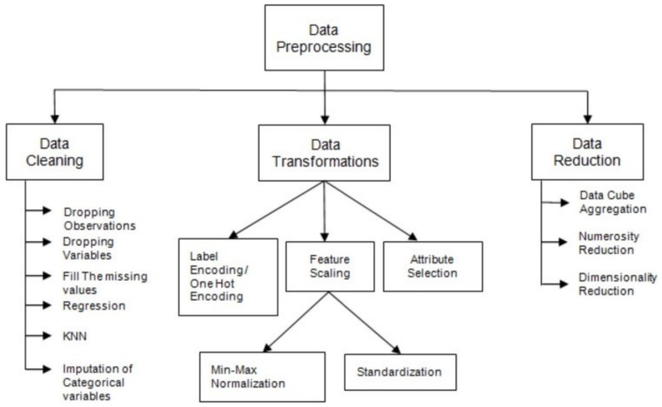
B. Data sets are accessible from making a caption generator

Information and knowledge are the basic keys the ideas going on in the human brain. It has been noticed that the rules and regulations that have a bulk of data and information in it are most likely to be unfollowed by the people. In earlier image caption generator methodology there were various datasets were used having large size or may be small size also, instance, Flickr8k, Flickr30k [17], MSCOCO, PASCAL, STAIR and AI Challenger Dataset. There are five images as reference in the database. To display the same image, there is a different grammar used so that the previous image can be displayed properly. For example, the Microsoft team created the Microsoft COCO Captions Database just to use the image captioning method to take generate the caption on daily basis and can be used to assign the tasks and the works, for example, classification and display, image recognition.

To build the required model of image captioning, we have used MSCOCO dataset having number of data/images equal to 82780.

C. Steps involved in data processing

Data purification: Raw data may miss data or have insignificant amounts of it. This can lead to inconsistencies in data. The data cleaning step handles all of these problems. All the values which are missing are rectified in accordance with the relevant terms and all non-essential amounts are deducted. Creating Test and Training Sets: This is one of the most important steps in data development. In this step we look at the data set in the training and setting of the test. For the training purpose a large portion of the data (70% -80%) is used and a small portion of the data (20% - 30%) is used for the experiment objective. Feature measurement: One variable dominates the other variable, if independent variables are measured at different scales. Feature measurement is therefore an important step in which all the values in the data are converted to a definite distance.



1. Flowchart

D. Model analysis

1) Supervised image captioning: For our paper, we will be using captions using CNN (Convolutional Neural Networks) and specialized RNN i.e. LSTM (Long Short- Term Memory). Image highlighting will be removed from inceptionv3/vgg16 model which is a CNN model based on image databases and thereafter feeds on the outstanding LSTM model which is responsible for producing the captions of a given image.

Convolutional Neural Network(CNN)
Convolutional Neural Networks are deep neural organizations that can deal with information in the form of input as a 2-D framework. We convert the images into a 2-D frame which are then fed into the CNN and then we can use it to successfully classify/identify the given image. In this way, CNN provides us an easy way to work with images. We primarily use CNN for image orders, image recognition, flight or Superman, and so on. It filters images from left to right and rips to extract highlights from the image and merges part of the group photos. It can deal with translated, uplifted, enhanced images and changes in context.

Recurrent Neural Network(RNN)
Intermittent Neural Network is an assertion of a feed-forward neural organization with memory within. The RNN is intermittent as it plays the same power in each data offering while the yield of current information is based on a single previous calculation. Following the harvest, it is repeated and sent back to the repetitive organization. By resolving the options, it looks at current data and yields obtained from previous information.
Long-Term Memory (LSTM)
Long Short-Term Memory (LSTM) networks is a type of RNN, it is an altered version of duplicate neural networks, which is adjusted in a way so that it can remember past data in memory easily. The RNN disappearance issue is resolved. The LSTM is improvised upon RNN to categorize, process and improvise a series of time given a period of undetermined duration. Back distribution is used to train the model. In the LSTM network, there are three gates:

- Input gate: This gate helps determining the part of input data which is to be converted to memory. The function sigmoid determines which values you must allow to exceed 0,1. and the performance of tanh gives weight to the transferred values that determine their value level from 1 to 1.

$$i_t = \sigma(W_i.[h_{t-1},\ x_t] + b_i) \tag{1}$$

$$C_t\ = tanh(W_C.[h_{t-1},\ x_t] + b_C) \tag{2}$$

- Forget gate: Find out what information will be lost on the block. Determined by the function sigmoid. It observes the previous situation (ht-1)and input text(Xt)and subtracts a number in between0(leave this)and1 (keep this) for each of the numbers in styleCt-1.

$$f_t = \sigma(W_f\ .[h_{t-1},\ x_t] + b_f) \tag{3}$$

- Output gate: The output of the LSTM structure is represented or provided by the this gate.

$$O_t\ = \sigma(W_o.[h_{t-1}\ ,\ x_t] + b_o) \tag{4}$$

$$h_t\ =\ O_t * tanh(C_t) \tag{5}$$

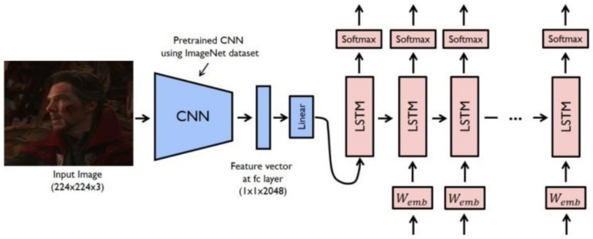


Figure 1. CNN-LSTM architecture for image captioning model

In this paper, the CNN model which is generally used for feature extraction of an image is replaced by VGG16/Inceptionv3 models as these models provide better accuracy and the image captions are predicted using beam search with different value of k and argmax search and then compared with each other for the accuracy and loss/validation loss. Flickr8k dataset and MSCOCO dataset is used here in this paper for experimentation.

- Approximately 1 hour is required for training of model/feature extraction for VGG16 model with Flickr8k dataset and 5-6 hours for MSCOCO dataset.
- Due to the fact that Inceptionv3 has very less parameter as compared to the VGG16 model so, it took 20 minutes for feature extraction with Flickr8k dataset and 2 hours 45 minutes for MSCOCO dataset using Inceptionv3 model.

2) *Unsupervised image captioning:* We have used the concept of machine translation using unsupervised technique for image captioning using unsupervised deep learning algorithm. We have considered out image as source language. We have used set of images I and sentences S^, and detector which detects visual concept. Ni is total numbers of images and Ns is total number of sentences. External corpus is used to get the sentence, they are unrelated to the image description.

The Model
The model is comprised of three parts , first one is image encoder, second is sentence generator and third is a discriminator for discriminating real sentences or produced by the generator.

Encoder
CNN model is used to provide the most important feature or pixels of an image as a result fim:

$$fim\ =\ CNN\ (I) \tag{6}$$

We used Inception-V4 for the purpose of encoder.

Generator
We have used LSTM for the purpose of generation of sentences. It takes the image representation and converts it into sentence in the language understandable by a human brain to describe the image. It provides us the probability distribution of all the words conditioned on the feature representation of image and the words that are already generated from our vocabulary.
The following equations are used to sample the generated word from the vocabulary.

$$\begin{gathered} x_{-1} = \mathrm{FC}(fim), \\ x_t\ = W_{e}{}_{s_t}, t \in\ \{0\ ...\mathrm{n}\ -1\}, \end{gathered} \tag{7}$$

$$, \qquad [p_{t+1}, h^g_{t+1}] = LSTM^{\,g}(x_t, h^g_t), \mathrm{t}\ \in \{-1\ ...\mathrm{n} - 1\}, \quad s_t \sim p_t, \mathrm{t}\ \in\ \{1\ ...\ \mathrm{n}\}$$

,

$$FC\ \ =\ Fully\ \ connected\ \ layer,$$

$\sim = operation\ \ used\ \ for\ \ sampling$

$n\ =\ generated\ \ sentence's\ \ length$

$W_e\ =\ \ word\ \ embedding\ \ \ matrix,\ x_t\ =\ input\ \ of\ \ LSTM$

$s_t\ =\ Representation\ \ of\ word\ generated\ \ by\ \ generator\ \ using\ \ one-hot$

$$h^g_t\ =\ hidden\ state\ of\ LSTM$$

$$h^g_t\ =\ hidden\ \ state\ \ of\ \ LSTM$$

$p_t\ =\ probability\ \ over\ \ the\ dictionary\ \ at\ the\ time\ step\ \ t$

$s_o\ =\ sentence\ \ starting\ \ point,\ \ s_n\ =\ \ sentence\ \ ending\ \ point$

s_t *is sampled from the probability distribution* p_t

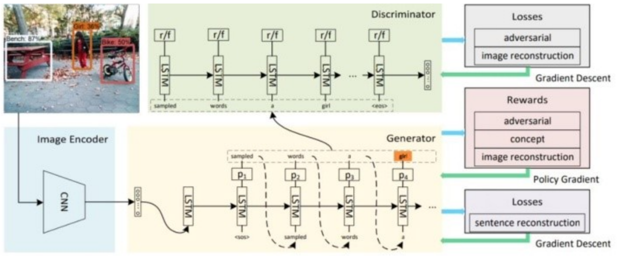


Figure 2. Model Architecture for unsupervised implementation of image caption generator.

Discriminator

We have used LSTM to implement the discriminator whose work is to discriminate the corpus contained real sentence from the sentence that the model generates.

$$[q_t, \square_t^d] = LSTM^d(x_t, \square_{t-1}^d), \mathrm{t} \in \{1 \dots \mathrm{n}\} \qquad (8)$$

Training

As this is not supervised training. So, there are three objectives of the training in this paper.

- **Adversarial caption generation**
  An adversarial text generation method [22] is used in this paper. The image feature is passed as an input to the generator which generates a sentence best suited tothe feature of image.

  The main work of generator is to generate sentences as

  real as possible so that it can fool the discriminator. To accomplish this, a reward mechanism is used in each time-step which is known as "adversarial reward". The value for the reward for the t-th generated word can be calculated as:

  $$r_t^{adv} = \log(q_t) \qquad (9)$$

  Adversarial loss for the discriminator is defined as:

  $$L_{adv} = -\left[\frac{1}{l}\sum_{t=1}^{l} \log(\hat{q_t}) + \frac{1}{n}\sum_{t=1}^{n} \log(1 - q_t)\right] \qquad (10)$$

- **Visual Concept Distillation** The adversarial type of reward only ensures generation of plausible sentences but they may be irrelevant to the image. So it is important to identify all the visual concepts present in the image. We have used a visual concept detector for this purpose. When a word generated by the model where the image is used to identify and detect the analogous visual concept, a reward named as concept reward is given to the word which is generated, where the confidence score of the visual concept corresponds to the reward value. The visual concept detector takes an Image I and yields a set of different concepts and their confidence scores.

  C=(c1,v1),...,(ci,vi),...,(cnc,vnc)
  C= Confidence Scores
  ci= visual concepts corresponding to ith value
  vi= confidence score of ith detected visual
  N_c =total visual concepts
  The t-th generated word is assigned a concept reward which is given by:

- **Bi-directional Image-Sentence Reconstruction**
  Limited Number of object concepts can be detected reliably by the existing visual concept detectors. It's important for our model that it has a better understanding of semantic concepts of the image. The images and sentences are put together into a common latent space for reconstructing each other.

Image Reconstruction

We have used the sentence generated by the generator reconstruction of image feature instead of constructing the whole image. So as it can be seen in the figure 5 discriminator is also an encoder for sentences. A layer that is fully connected is arranged on the top of the discriminator so that the last hidden state can be papered to the latent space that is common to both sentences and images:

$$x' = FC(h_n^d) \qquad (11)$$

For the purpose of training the discriminator an additional image reconstruction loss is defined:

$$L_{im} = ||x_{-1} - x'||_2^2 \qquad (12)$$

$$r_t^{im} = -L_{im} \qquad (13)$$

r_t^{im} = *Reward to the generator for reconstructing* the image

L_{im} = *Reconstruction error*

Sentence Reconstruction

The derived representation can be used by the generator to reconstruct the sentences. It is also a type of sentence denoising auto-encoder [23]. The images and sentences are aligned by it in the latent space, and by having the common space for the representation of an image, the learning of the decoding of sentence takes place. The cross-entropy loss is defined as:

$$L_{sen} = -\sum_{t=1}^{l} \log(p(s_t = \hat{s_t} | \hat{s_1}, \dots, \hat{s_{t-1}})) \qquad (14)$$

Integration

In this paper policy gradient [21] is used for training the generator, where the gradients are estimated with respect to all the trainable parameters. The gradients for the generator are provided by the sentence reconstruction loss using back-propagation. For updating the generator both of the gradients are used. The gradient is:

$$\nabla_\theta \mathscr{L}(\theta) = -\mathbb{E}\left[\sum_{t=1}^{n}\left(\sum_{s=t}^{n} \gamma^s \left(r_s^{adv} + \lambda_c r_s^c\right) + \lambda_{im} r_s^{im} - b_t\right) \nabla_\theta \log\left(s_t^T p_t\right)\right] + \lambda_{sen} \nabla_\theta \mathscr{L}_{sen}(\theta) \qquad (15)$$

The combination of image reconstruction and adversarial losses provide the updating required with the parameters with the help of gradient descent process within the discriminator.

$$L_D = L_{adv} + \lambda_{im} L_{im} \qquad (16)$$

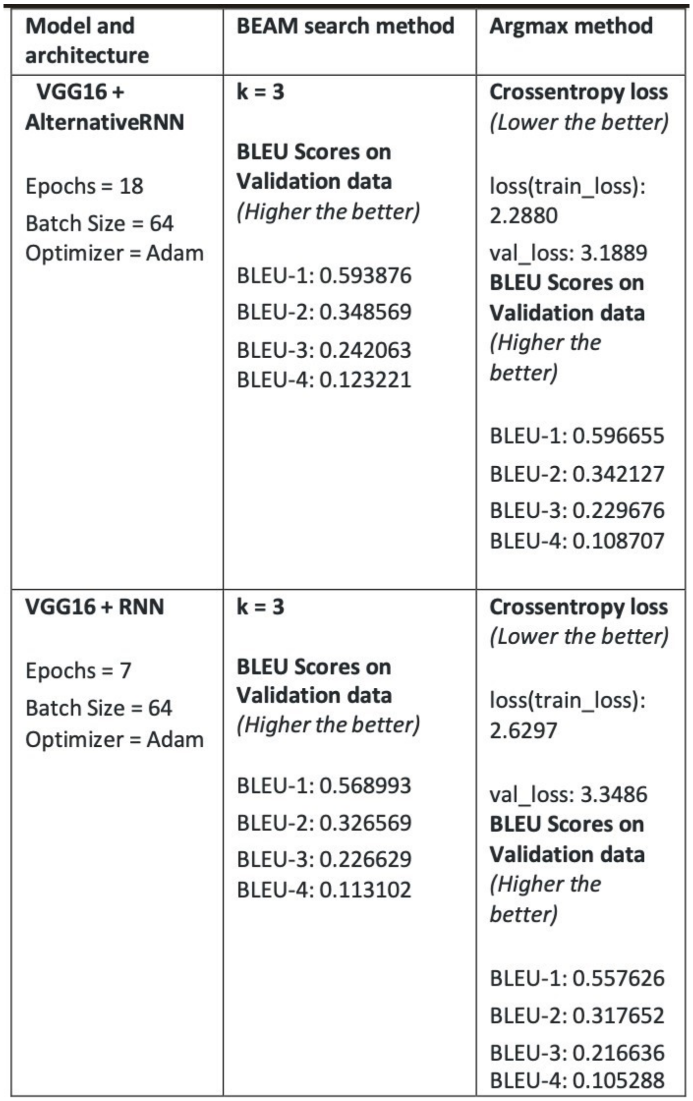
IV. Results and Discussion

In this paper, as pre-trained models has been used for the purpose of feature extraction which is then feed as the input vector to the RNN/LSTM model/architecture for the final image captioning. The model VGG16 has larger number of parameters as compared to the Inceptionv3 model that is why, the later model takes lesser time for the feature extraction/training of the input dataset.

For image caption prediction, two methods have been used in this paper, beam search and argmax search and the results are compared with respect to the different pre-trained models using tables. The criteria for the comparison is taken to be the loss/validation loss value instead of accuracy value and the standard metric for comparison used here is BLEU score.

Initialization A pipeline is proposed for the pre-training of the discriminator and the generator. For each of the training image we have generated a pseudo caption, then image captioning model is initialized by the pseudo image-caption pairs. A concept dictionary has been built which contains all the object classes present in the dataset OpenImages [20].

Here, sentence corpus is required and used to provide the required training of a concept to the sentence model. The concept of each image provided is effectively detected using existing visual concept detector. Then the training of generator has been done with the help of standardized deep learning algorithm which includes pseudo pairs of image caption[1].



Model and architecture	**BEAM search method**	**Argmax method**

VGG16 + AlternativeRNN
Epochs = 18
Batch Size = 64
Optimizer = Adam
	k = 3
	BLEU Scores on Validation data *(Higher the better)*
	BLEU-1: 0.593876
	BLEU-2: 0.348569
	BLEU-3: 0.242063
	BLEU-4: 0.123221
		Crossentropy loss *(Lower the better)*
		loss(train_loss): 2.2880
		val_loss: 3.1889
		BLEU Scores on Validation data *(Higher the better)*
		BLEU-1: 0.596655
		BLEU-2: 0.342127
		BLEU-3: 0.229676
		BLEU-4: 0.108707

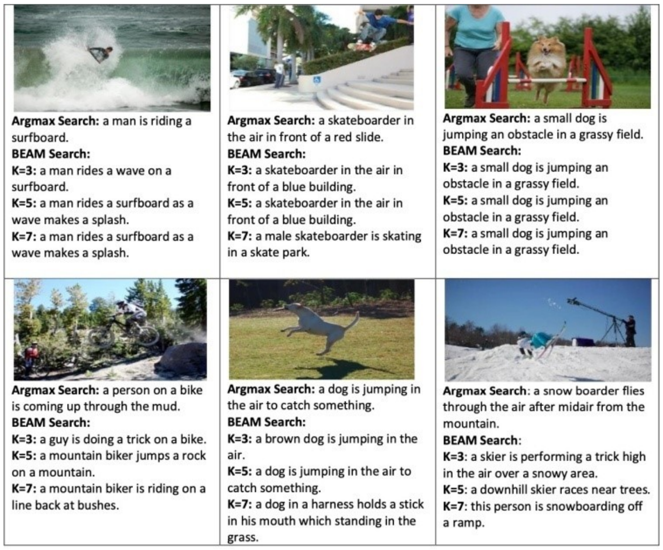
VGG16 + RNN
Epochs = 7
Batch Size = 64
Optimizer = Adam
	k = 3
	BLEU Scores on Validation data *(Higher the better)*
	BLEU-1: 0.568993
	BLEU-2: 0.326569
	BLEU-3: 0.226629
	BLEU-4: 0.113102
		Crossentropy loss *(Lower the better)*
		loss(train_loss): 2.6297
		val_loss: 3.3486
		BLEU Scores on Validation data *(Higher the better)*
		BLEU-1: 0.557626
		BLEU-2: 0.317652
		BLEU-3: 0.216636
		BLEU-4: 0.105288

Table 1. BLEU scores based comparison for VGG16 model.

Model and architecture	**BEAM search method**	**Argmax method**

InceptionV3 + AlternativeRNN
Epochs = 20
Batch Size = 64
Optimizer = Adam
	k = 3
	BLEU Scores on Validation data
	(Higher the better)
	BLEU-1: 0.606086
	BLEU-2: 0.359171
	BLEU-3: 0.249124
	BLEU-4: 0.126599
		Crossentropy loss
		(Lower the better)
		loss(train_loss): 2.4050
		val_loss: 3.0527
		BLEU Scores on Validation data
		(Higher the better)
		BLEU-1: 0.596818
		BLEU-2: 0.356009
		BLEU-3: 0.252489
		BLEU-4: 0.129536

InceptionV3 + RNN
Epochs = 11
Batch Size = 64
Optimizer = Adam
	k = 3
	BLEU Scores on Validation data
	(Higher the better)
	BLEU-1: 0.605097
	BLEU-2: 0.356094
	BLEU-3: 0.251132
	BLEU-4: 0.129900
		Crossentropy loss
		(Lower the better)
		loss(train_loss): 2.5254
		val_loss: 3.1769
		BLEU Scores on Validation data
		(Higher the better)
		BLEU-1: 0.601791
		BLEU-2: 0.344289
		BLEU-3: 0.230025
		BLEU-4: 0.108898



Argmax Search: a man is riding a surfboard.
BEAM Search:
K=3: a man rides a wave on a surfboard.
K=5: a man rides a surfboard as a wave makes a splash.
K=7: a man rides a surfboard as a wave makes a splash.

Argmax Search: a skateboarder in the air in front of a red slide.
BEAM Search:
K=3: a skateboarder in the air in front of a blue building.
K=5: a skateboarder in the air in front of a blue building.
K=7: a male skateboarder is skating in a skate park.

Argmax Search: a small dog is jumping an obstacle in a grassy field.
BEAM Search:
K=3: a small dog is jumping an obstacle in a grassy field.
K=5: a small dog is jumping an obstacle in a grassy field.
K=7: a small dog is jumping an obstacle in a grassy field.

Argmax Search: a person on a bike is coming up through the mud.
BEAM Search:
K=3: a guy is doing a trick on a bike.
K=5: a mountain biker jumps a rock on a mountain.
K=7: a mountain biker is riding on a line back at bushes.

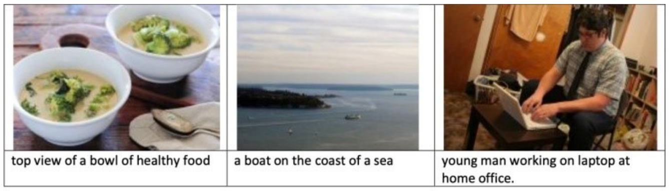
Argmax Search: a dog is jumping in the air to catch something.
BEAM Search:
K=3: a brown dog is jumping in the air.
K=5: a dog is jumping in the air to catch something.
K=7: a dog in a harness holds a stick in his mouth which standing in the grass.

Argmax Search: a snow boarder flies through the air after midair from the mountain.
BEAM Search:
K=3: a skier is performing a trick high in the air over a snowy area.
K=5: a downhill skier races near trees.
K=7: this person is snowboarding off a ramp.

Figure 3. Comparing captions for images for different value of k for beam search method and argmax method used.

Here, in figure 3 we can observe that beam search method for prediction of caption of an image provide more accuracy and preciseness than the argmax method and within the beam search method, higher the value of k, better the results will be obtained. Above we discussed the result and discussion part of the supervised image captioning now in the later part we are going to discuss the outcome of the image captioning model when implemented through unsupervised manner.

- Shutterstock1 is used here to gather the image description which is eventually used to collect the corpus of sentence. Here, visual concept detector is used by using object detection model [27] which is properly trained on OpenImages [20].
- The model built for image caption generator using unsupervised algorithm provide enhanced and better results for accuracy with 28.9% if it is measured with respect to CIDEr score.
- The space dimension for shared latent and the dimensions which are available for LSTM are taken to be 512 which is a fixed value and the to achieve the approximately same level and scale of weighting parameters with respect to different rewards, the value of sen, im, c, are taken to be 1, 0.2,10 and 0.9 respectively.
- The training of the model or the dataset is done using Adam optimizer[26] and it has the learning rate of 10-4. In the initial phase, the optimization/minimization of the loss or we can say it a cross entropy loss is achieved by taking into consideration the learning rate to be 10-3. Consequently, during the testing of the model the beam search method is opted with the value of k taken to be 3.
- We have used both supervised learning and unsupervised learning to build the model of image caption generator and results were very noticeable and promising but as we have used sentence corpus not image sentence pairs in unsupervised learning so the captions are more diversified. Our vocabulary is huge in unsupervised leaning. Captions generated in unsupervised learning are more plausible and can be seen in figure 4.



top view of a bowl of healthy food

a boat on the coast of a sea

young man working on laptop at home office.

Figure 4. Qualitative image caption analysis for unsupervised manner for image caption generator.

V. CONCLUSION

As various earlier studies in the area/field of image captioning model has been done providing the state-of-art model which primarily generates captions for image having the architecture of encoder-decoder (i.e., CNN-RNN pair). Here, enhancement in accuracy and reduction in loss/validation loss is achieved as some of the already trained imagenet models are taken into consideration for the feature extraction phase of the model building. The comparison among results for these models is done for the accuracy and then different methods used for the prediction of captions for input image are also compared. It is observed that the model is more improved with respect to the memory issue and accuracy if batch size of the dataset during the training of the model is taken to be of higher value and the value of k is higher in beam search method. The unsupervised method for image captioning is also experimented and model is built for it which shows improvement in the performance and accuracy of the model for image caption generator. As corpus which is an image describing content consist of over two million sentences gathered from Shutterstock used efficiently for the purpose of implementing unsupervised algorithm for image captioning that subsequently lead to the promising, more efficient and accurate results even if the image sentence label is not provided.

REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, ”Show and tell: A neural image caption generator,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp.3156-3164, doi: 10.1109/CVPR.2015.7298935.

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

[3] A. Aker and R. Gaizauskas. Generating image descriptions using dependency relational patterns. In ACL, 2010.

[4] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. C. Berg, K. Yamaguchi, T. L. Berg, K. Stratos, and H. D. III. Midge: Generating image descriptions from computer vision detections. In EACL, 2012.

[5] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In ACL, 2012.

[6] P. Kuznetsova, V. Ordonez, T. Berg, and Y. Choi. Treetalk: Composition and compression of trees for image descriptions. ACL, 2(10), 2014.

[7] D. Elliott and F. Keller. Image description using visual dependency representations. In EMNLP, 2013.

[8] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In CVPR, 2011.

[9] R. Kiros and R. Z. R. Salakhutdinov. Multimodal neural language models. In NIPS Deep Learning Workshop, 2013.

[10] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computation, 9(8), 1997.

[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In arXiv:1502.03167, 2015.

[12] Andrej Karpathy, Li Fei Fei (2015) Deep Visual-Semantic Alignmentsfor Generating Image Descriptions.IEEE Transactions on Pattern Analysis and Machine Intelligence(April 2017), vol 39, issue 4:664 –676

[13] T.-Y. Lin, et al (2014) Microsoft COCO: Common objects in context.arXiv:1405.0312

[14] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Rossand Vaibhava Goel (2016) Self-critical Sequence Training for ImageCaptioning.in arXiv:1612.00563

[15] Ronald J. Williams (1992) Simple statistical gradient-following algorithms for connectionist reinforcement learning.In Machine Learning,pages 229–256, 1992.

[16] Marc’ Aurelio Ranzato, Sumit Chopra, Michael Auli, and WojciechZaremba. (2015) Sequence level training with recurrent neural networks.ICLR, 2015.

[17] P. Mathur, A. Gill, A. Yadav, A. Mishra and N. K. Bansode, ”Camera2Caption: A real-time image caption generator,” 2017 International Conference on Computational Intelligence in Data Science(ICCIDS), Chennai, India, 2017, pp. 1-6, doi: 10.1109/ICCIDS.2017.8272660.

[18] Haoran Wang, Yue Zhang, Xiaosheng Yu, “ An Overview of Image Caption Generation Methods”, Computational Intelligence and Neuroscience, vol. 2020, Article ID 3062706, 13 pages, 2020. https://doi.org/10.1155/2020/3062706.

[19] T.-Y. Lin, et al (2014) Microsoft COCO: Common objects in context.arXiv:1405.0312.

[20] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu- El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi- label and multi-class image classification. Dataset available from https://storage.googleapis.com/openimages/web/index.html, 2017.

[21] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In NIPS, 2000.

[22] William Fedus, Ian Goodfellow, and Andrew M Dai. Maskgan: Better text generation via filling in the . In ICLR, 2018.

[23] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In ICML, 2008.

[24] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data,modelsand evaluation metrics. JAIR, 47, 2013.

[25] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internalcovariate shift. In arXiv:1502.03167, 2015.

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[27] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In CVPR, 2017.

[28] Ranganathan, G. "Real Life Human Movement Realization in Multimodal Group Communication Using Depth Map Information and Machine Learning." Journal of Innovative Image Processing (JIIP) 2, no. 02 (2020): 93-101.

[29] Bindhu, V., and Villankurichi Saravanampatti PO. "Semi-Automated Segmentation Scheme for Computerized Axial Tomography Images of Esophageal Tumors." Journal of Innovative Image Processing (JIIP) 2, no. 02 (2020): 110-120.

[30] S. Han and H. Choi, "Domain-Specific Image Caption Generator with Semantic Ontology," *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Busan, Korea (South), 2020, pp. 526-530, doi: 10.1109/BigComp48618.2020.00-12.

[31] Vijayakumar, T., and R. Vinothkanna. "Retrieval of complex images using visual saliency guided cognitive classification." J Innov Image Process (JIIP) 2, no. 02 (2020): 102-109.

[32] N. K. Kumar, D. Vigneswari, A. Mohan, K. Laxman and J. Yuvaraj, "Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach," *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, Coimbatore, India, 2019, pp. 107-109, doi: 10.1109/ICACCS.2019.8728516.