ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1602

A Comprehensive Guide to Deep Neural Network-Based Image Captions

Pritesh Pandey (Ph.D Scholar)
Gujarat Technological Univeristiy

Dr. Keyur N. Brahmbhatt
Birla Vishvakarma Mahavidyalaya Engineering College

Abstract:- A sore subject for understanding an Image is Image captioning. It is the amalgamation of two key components in. look and language expression which refers to 'NLP (Natural Language Processing)' & 'Machine Vision' which are considered the most prominent areas of computing. Image captioning approach has advanced rapidly because of the events of higher labeling information and deep neural network. The image captioning techniques and enhancement supported deep neural networks are presented along with the features of specific approaches in this study. The retrieval-based method is the foremost image captioning technique premised on deep neural networks. The recovery technique takes advantage of a looking approach to seek out an applicable image specification. The •template based' approach segregates the image tagging technique to item recognition along with statements procreation. For Image Captioning the end to end learning based techniques have been substantiated remarkably effective. Renewed dexterous and facile statements can be procreated by end-to-end learning. In course of the study, approaches related to Image Captioning are examined completely along with the discussion of other remaining challenges.

I. INTRODUCTION

Every day, we have a tendency to encounter an outsized variety of pictures from numerous sources like the net, media sections, documentation pattern and commercials. Such inceptions comprehend pictures which audience ought In decipher by them. The majority of pictures don't posses an outline, however the audience will for the most part perceive them while not their elaborate labels. Nevertheless if automatic captioning of image is needed by audience (humans), than machines have to translate some styles of Image Caption

There is a great advancement of deep neural network approaches in past decades and is of a great success when used with the applications for instance MT, sonnet identification, and machine vision. Image captioning comprised of equipping a concise and elliptical explanation for an image in lingual connection and is currently consummate by methods that utilize s an association of lingual communication process (NLP), pc vision (CV), and machine learning techniques. This work will be demonstrated with the help of a machine that states a picture and translates it into a statement or a section in accordance with its apprehension. This accomplice activity is naive for humans, together with tiny kids, nevertheless complicated for a machine. Captioning not solely demands the utilization of a prototype to distinguish the section of a image and comprehend their association nevertheless in addition to, it have to categorize the fundamental idea of this purpose in lingual transmission, seize the lingual knowledge of a picture, and produce human- legible statements. Image captioning has substantial notifications, specifically in AI, for the reason that it will commend instruments "see" the substance of an image, encourages machine intellect, and employs th at intellect well to look engines.

Understanding a picture for the most part is influenced by on obtain image alternatives. The methods for use therein will be broadly segregated into two groups: (I) Approach based on former machine learning and (2) Approach based on Deep machine learning.

Traditional options like SIFT (Scale- Invariant Feature rework), the bar graph of familiarized Gradients (HOG), native Binary Patterns (LBP) [40], and a mixture of such alternatives area unit wide are used in ancient machine learning. To categorize AN object the alternatives area unit is retrieved from input data set. To categorize an object, they're provided in a classifier like Support Vector Machines (SVM) [26]. As traditional alternative area unit work specified, retrieving alternatives by mean of an enormous and different range of knowledge shall not be feasible. Furthermore, universe in formation like images and video area unit progressed and has completely various lingual explanations.

Across the other side, in approaches based on deep machine learning, alternative region unit erudite mechanically from training information and that they will control an large and numerous category of pictures and videos as an example, CNN (Convolutional Neural Networks) [38] area unit wide used for characteristics SoftMax is utilized during a classifier like To get labels CNN is primarily accompanied by continual Neural Networks.

In the last five years, An outsized variety of sections are revealed on image captioning along with deep machine learning being prominently utilized. The intricacies and difficulties of image captioning will be managed well enough by Deep learning algorithms. As of today, solely three study article s [23, 24, 37] are identified over this review matter. Though the articles offered a decorous literature study of image captioning, it might merely mantle couple of articles on deep learning as a result majority of them was revealed once the survey articles. These survey articles principally mentioned template based mostly, retrieval based mostly, and a really some unique image.

Captions based on deep learning procreating prototypes. Nevertheless, an enormous diversity of tasks is employed ‹xi image tagging which is based on steep learning. Furthermore, learning-based image captioning and remarkable analysis space has been created by the supply of enormous and new collection of data. To produce an short version of the literature, we have a tendency to gift a survey principally that specializes in image captioning based ‹xi deep learning studies.

The primary objective of study is to procreate an extensive review of deep learning-based image captioning. Initially, we should have a propensity to 'aggregate the prevalent inn age caption ing studies into three main classes:

(I) Image Captioning based on template, (2) Image Captioning based on retrieval, and (3) Unique image caption procreation. The classes are a bit mentioned shortly in Section a pair of. Most of the deep learning based mostly image captioning strategies comprise the class of unique caption procreation. Thus, we have a propensity to focus solely on unique caption procreation through deep learning. Secondly, we have a propensity to a negate the strategies of deep learning based image cap g into completely distinct categorize specifically into (1) Supervised learning, (2) Alternative deep learning, (3) Based on Visual Space, (4) Based on Multimodal, (5) Based on Encoder-Decoder Architecture, (6) Based on integrative Architecture, (7) Den se captioning, (8) Whole scene- based mostly, (9) Attention-Based, (10) linguistics concept-based, (11) Long short-run Memory (LSTM) [54] Based on language mrx4el, (12) Based on others language model, (13) artificial tagging, and (14) Image captioning based on Novel object.

II. IMAGE CAPTIONING

Three strategies which can be used for image captioning supported by creep neural network approaches are: method based on retrieval, methods based on template and based ‹xi end-to-end learning approaches. Based on Retrieval image captioning approaches, which area unit is currently seldom applied is discussed in Section 1. Template-based strategies area unit mentioned in Section 2. different techniques area unit and strategies based on end -to-end learning are discussed in Section 3.

A. Section-1 Retrival Method

Recovery or retrieval method generates the use of some excerpts to obtain deep sensory neural networks and deep auto encrxlers with alternating words of the image. Accordingly, a good and accurate picture pursuit approach is designed. Eventually, this prototype employs various resources based on Internet to examine the lingual knowledge of pictures further, and then uses an equivalent picture search approach to conclude the lingual of latest images in conduit along with commonalities. In summary, equally captioned pictures square measure 1st retrieved from an oversized dataset then adapted to suit the unique pictures [1]. Similar ways usually includes abstraction as associate degree intermediary stage to undermine or take away information's of a caption affiliated solely with the extracted picture. This technique simply receives lingual knowledge however needs the coaching datasets consisting of all types of intents. It is a higher reliance on the seed images, and that is not intelligent at exploring terms external the instruction information. Accordingly, it will be executed right solely with decent excellence coaching knowledge. Unique recovery primarily founded ways utilize neural networks to make the connection among pictures and topic by a standard vector mapping [2-4]. bound approaches based on retrieval preference picture alternatives and utilized analogy metrics [5, 6]. Such ways will get grammatical captions however cannot describe the items that are not within the training dataset.

B. Section-2 Template Method

Scholars have searched -after unique means that to appreciate image captioning thanks to the retrieval approaches barriers. The majority of the scholars take into account that such work may be at the start segregated into two partial operation. The principal subtask implies target identification and categorization exploitation computer vision approaches [7]. The secondly partial operation is that the conveying of discovered articles by statements exploitation language prototypes (LM s) in step with the properties of those items and also the connection within the items and also the environment. Thus, scholars use the prototype approach to amalgamate the results of the two partial operations to understand image captioning. Such techniques use optical object sensor to acknowledge and examine objects in pictures so as for searching out group of terms usually enclosed within the captions. The identified terms area unit then channeled to the model, several connected terms or expressions to come up with a complete statement [8]. every unit of those methods can be troubleshoot separately.

The procreation of picture representation from the collection of doubtless words utilization a good luminous flux unit may be thought about as an enhancement technique. The work targets to search out the statements that the majority doubtless comprehends the identified texts to explain the picture. an object detector was used by Farhadi et al. to deduct the triplet of scene pieces so renewed it into a statement in stages with definite patterns. Kulkar ni anci Li [8] conducted identical tasks, that amalgamate the identified items to come up with a ultimate statement. For the language, statement creation is asserted on the expression connected with items and their connections. Kulkarni et al. Make use of a posh model technique to sight items and create a statement. Scholars coming back from Microsoft projected an equivalent technique [9]. The technique 1st uses several example educations (MIL) to coach visible sensors to get terms regarding the pictures. A applied mathematics prototype is then coached In come up with the outline. Lastly, a reciprocation prototype is employed to judge and Gracie a group of statements with higher opportunity. The method is demonstrated in Fig. 1 .

Image explanation includes adjective, verbs and nouns. Consequently, the tagging captions prepare the label creator could be a rational alternative. Detecting devices will discover same epitome terms, like "beautiful" and "riding." These terms could have a substantial connection with particular visible models, like a bicycle or horse riding. Second, photos continuously includes info that demonstrates the logic of individuals. coaching luminous flux unit by image captioning will seize this logic right. for instance, LM earns with logic that someone sitting on a chair is healthier than the person who is standing on the chair.

C. Section-3 End-to-End Learning Method

End-to-end learning strategies will notice perception and Interpretation of pictures by a right of way prototype. Each criterion will be erudite immediately by the mean of training . Such technique was motivated by the current advancement of MT and order-to-order prototypes. kind of like the machine interpretation method [10-11], the interpretation prototype resourced a statement of language so interprets it into associate purpose language. within the recent years, MT was accomplished by a sequence of individual works, like interpreting, interconnecting, and rearrangement terms. Currently, with the advancement of deep neural network-based MT prototype, RNN has been forecasted to achieve the work [10, 12]. Image captioning orders to transform connected entry picture into a statement, that is analogous to MT. Google utilizes associate method that's supported a mix of RNN and CNN [l3]. Deep CNN (DCNN) is employed as associate encrypt or to browse and remodel entered pictures into vectors for perception of the picture, that is analogous to what the RNN in MT prototype will. To interpret the entered picture to a statement the interpreter RNN utilizes the vector as connection to primary value of the hidden layer.

Image captioning strategies based on template would like term identification, statements procreation by lumen, and grading of statements. The end-to-end training prototype amalgamate DCNN along with RNN, along with utilization of the image and respective equivalent tags to coach the combine prototype [I 4] by optimizing the possibility of the created sentence S = (S I , S2, . . .) cherish the image I .. Such methods with success increase the capabilities of image tagging (captioning) by pc to a brand-new dimension. Various prior tasks transmitted what was included in an image or just delineated pictures by the language emerging under the training information bolt, and that they couldn't supply adequate explanation thanks to the ignored connections within the items and setting. On the premise of comprehending of the internetworking of items and setting in the pictures the specified system attains a new phase and might mechanically create an correct explanation for the pictures. This open system will procreate a additional smooth tongue to explain pictures.

The prototype utilizes Deep Convolutional Neural Network [16] as encrypting device to arrange picture s of RNN to rewrite the outcome of CNN and procreate statements. Neural Network was used by Kiros and Salakhutdinov to forecast terms in accordance with pictures and terms procreated antecedently. To arrange visible entry data directly Mao et al. [l7] produced multimodal repeated neural networks, consequently facultative the RNN to register the identified items. The two greatest crucial discoveries in the machine vision and true process are DCNN and LSTM. LSTM was utilized by Vinyals et al. [l3] and Donahue et al. [18] in their prototype. Kiros et al. [19] combined both the modern approaches for image captioning. The top most layer neuron outcome of DCNN will depict the high-level lingual knowledge of pictures, that could be utilized because the input data of LSTM to get image specification. Such a method amalgamates two modern approaches. Each are applied wherefore they are optimal at and accomplished along with end-to-end learning.

- *Supervised Learning Vs Alternative Deep Learning*

Coaching knowledge Keep Company with desired output referred to as label in supervised learning. Unattended training, on the opposite assist, deals with untagged instructions. Generative Adversarial Networks (GANs) [33] are a sort of unattended education approach. One more approach of machine learning is Reinforcement learning where the objective of an agent is to get knowledge and/or labels through investigation and a current indication. Reinforcement learning and GAN are utilized by image Captioning approaches. Such ways sit within the class of "Other Deep Learning".

- *Image Captioning Based on Supervised Learning.*

In image categorization [34, 36, 45], object detection [31, 32, 43], and property training [30] networks based on Supervised learning have been utilized over the years with success. This advancement makes scholars fascinated by victimization of them in automated image captioning [39, 46]. An outside range of supervised learning-based image captioning has been known in this study. We tend to categorize them into completely diverse classes: (i) Architecture based on Encoder-Decoder, (ii) Based on integrative design, (iii) Based on Attention, (iv) Based on linguistic concept, (v) Based on artificial captions, (vi) Based on Unique object, and (vii) Based on Dense image captioning.

- *Alternative Deep Learning*

In our daily life, knowledge is expanding with unlabeled knowledge as a result of its typically unrealistic to preciously interpret knowledge. As a result, currently, scholars are concentrating additional on reinforcement learning and unaccompanied image captioning techniques based on learning.

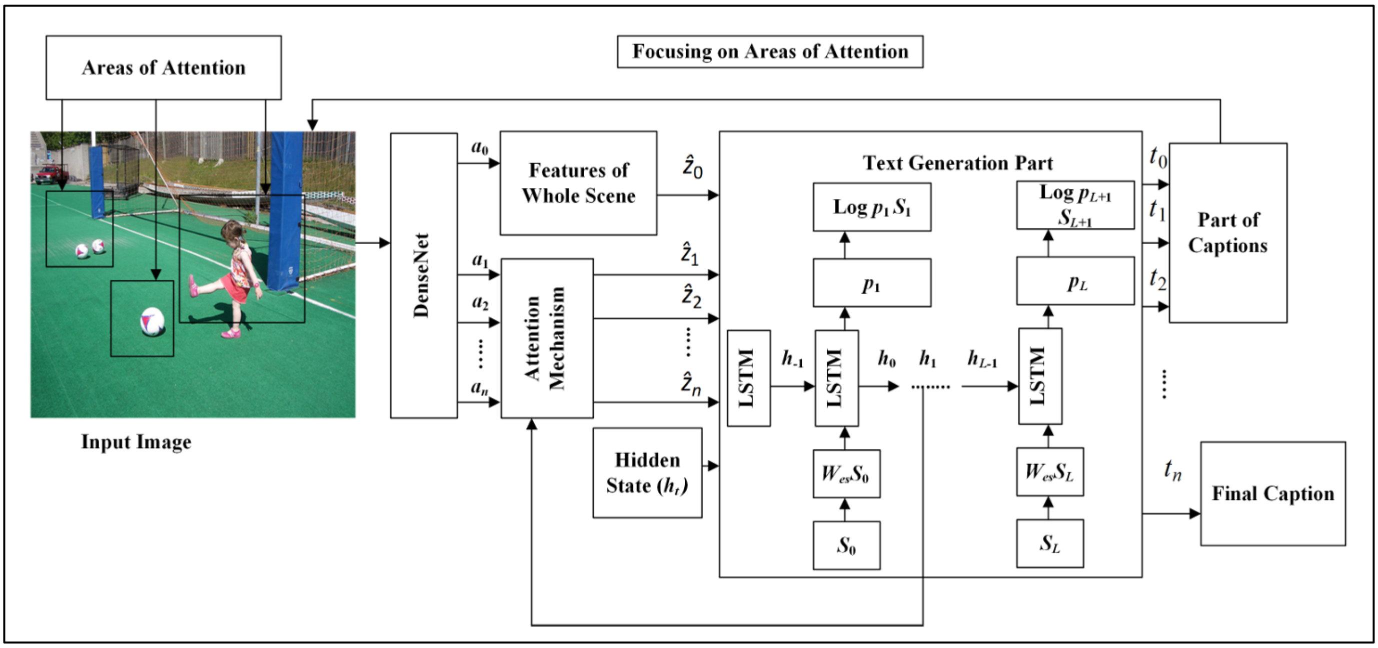


Fig. 1: Alternative Deep Learning Algorithm Framework

An action was selected by a consolidation training agent, acquires recompense worth, and shifts to a substitute state. The agent makes an attempt to pick out the operation with the expectancy of getting a most semi-permanent award. It desires constant state and operation data, to produce the assurances of a price operate. Ancient reinforcement training advances face a variety of constraints like the shortage of assurances of a price operate and unsure data. Procedure gradient ways are a sort of learning that may opt for a particular strategy during particular operation victimization gradient descent and improvement approaches. Domain data will be incorporated by the policy for the operation that assures the union. As a result, policy gradient ways need lesser standards than function under most techniques.

Variants of image encoders are utilized to retrieve picture alternatives by current image captioning based on deep leaning. To obtain captions, alternatives are supplied into the language decoder based on neural network. The ways contain 2 major problems: (i) they are skilled victimization most likely - hood evaluation (ii) back-propagation [41] techniques. For such cases, successive terms are anticipated given the picture and everyone the antecedent produced ground-truth terms. Consequently, the procreated tags seem like ground-truth tags. The development is termed revelation partiality downside.

Estimated metrics at verifying duration are non-distinguishable. Theoretically, sequential prototype for an image captioning needs to be prepared to prevent disclosure bias and immediately improve result-matrix for the verification duration. Critic is often utilized in calculating the anticipated future awards to coach the actor (captioning policy network) in reinforcement learning rule. Image tagging based on Reinforcement learning ways model and successive token from the model supported the awards which are accepted in every state. The gradient will be maximized by the Policy gradient ways in reinforcement learning so as to prognosticate the additional semi- permanent awards. Consequently, it will resolve the non- distinguishable downside of review metrics.

The approaches for this categorization pursue the subsequent stages:

- The merged network of RNN and CNN mostly generate captions.
- An additional network based on CNN-RNN estimates the tag s and transmits response to the basic network to acquire main quality tags.

A prototype of a technique of this class is demonstrated in Figure 3. Most methods based on GANs will learn deeper options from unlisted knowledge. They succeed this representation in implementing a competing method between combinations of networks: generators and anyone. GAN has previously been utilized with success in a plethora applications, as well as image captioning [28, 44], image to image translation [35], text to image synthesis [25, 42] and Text generation [29, 47]].

Two problems with GAN are: Firstly, GANs would job well in creating natural pictures from actual pictures since real- valued knowledge is planned as a result of GAN s. However, the textual process is predicted at different information. For that reason, such operations are non-separable, creating it complex to directly use backpropagation. Work in a constant amount to formulate the policy so that the gradients can be propagated back. Second, judges face problems in error propagation for missing gradients and series production. It wishes for a possible upcoming reward for each fractional explanation. The Monte Carlo rollout is employed to calculate the value of this future reward. GANs are primarily based mostly image captioning methods that will generate different sets of image captions, distinguishing traditional deep precipitation networks and deep perennial network-based models. Dai et al. [2 planned] additionally a GAN-based mostly image caption technique was plaid. However, they are not thinking of multiple tags for an image. Shetty et al. [4] introduced a replacement GAN-based mostly image caption technique. The technique will create numerous tags for an image and will show spectacular growth in creating multiple label s. GANs have restrictions for back propagating different knowledge. Gumbel sampler is employed to beat knowledge separately. The two main elements of this regressive network are generators and also. Throughout coaching, the generator learns the disadvantages provided by an individual rather than learning from express sources. One has a true knowledge allotment and may differentiate among generator-created examples and correct knowledge samples. Thus, this enables the network to detect multiple knowledge distributions. In addition, the network classifies the produced caption set as both actual and accurate. So, it will procedure tags just as a human produces one.

➢ *Captions for the Total Scene Vs Dense Captioning*

All fields of view captions are procreated in dense captioning. Alternative techniques create labels for the total view.

Dense Captioning. Prior image captioning methods will procreate only one tag for an image. Different areas of the image are used to obtain data of different items. Nevertheless, region-wise captions are not produced by these methods.

Johnson et al. [50] planned a picture captioning technique referred to Dense Cap. These approaches localize all the prominent areas of a picture, so it creates specifications for those areas.

➢ *A Standard Approach of this Section Consists Subsequent Sages:*

- Region projection are procreated for the different regions of the given image.
- To obtain region-based image options CNN is selected.
- The outcome of stride a pair of is utilized by a lingual prototype to get captions for each area

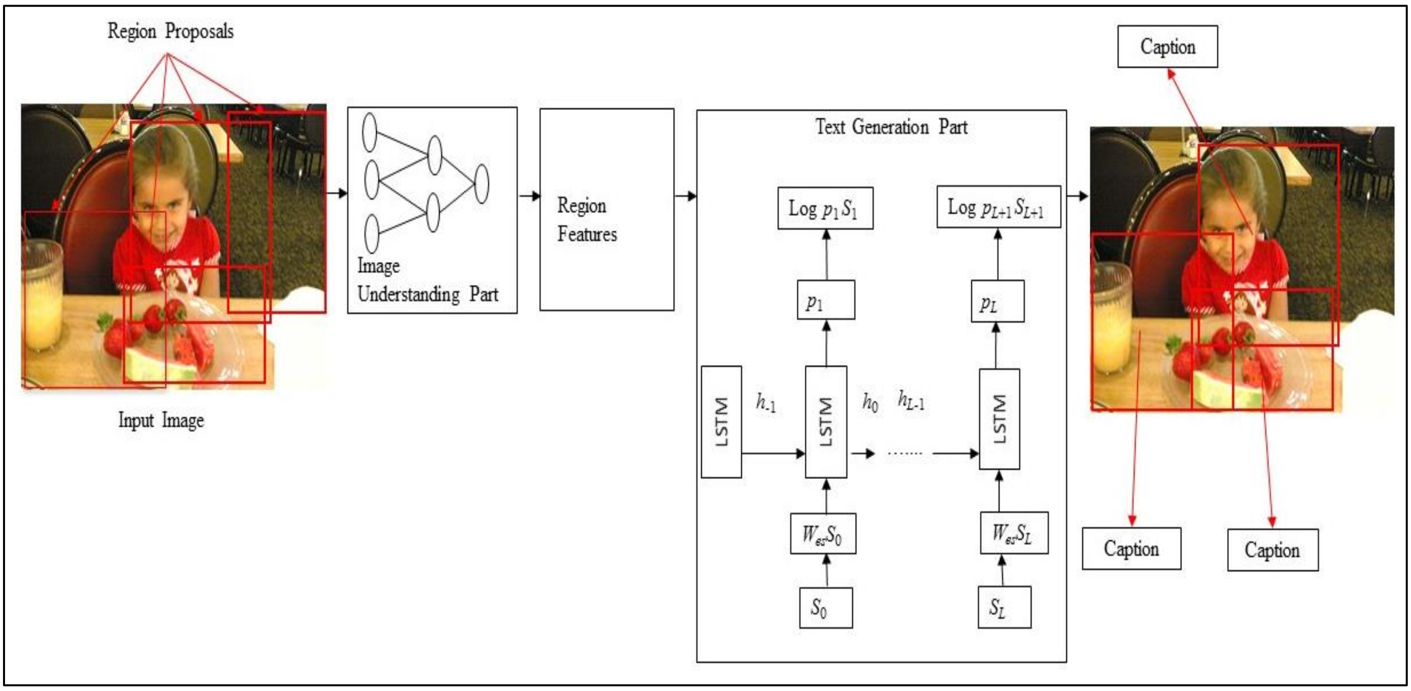


Fig 2: The layout of a common substantial feature technique

Those dense captioning[50] adduces an altogether assertive localization specification, which includes a tenacious network, a substantial localization layer, and a Long Short term Merory[49] Language protype. The substantial localization stack advances a picture with distinguished, economical passing play that presumes a collection of area of interest in the picture. Thus it does not require any external field resolutions to accelerate Region Convolutional neural network or a complete network. Image captioning techniques based on neural network-directly simplify task directly to eliminate. These methods are almost like neural MT based on encoder-decoder framework. During this network, CNN is world image options are extracted from hidden activities, so they are provided as an input to the ANM LTM to obtain a succession of words. Proposal network of RCNN: The working rule of the localization stack task is expounded with the task of quicker Region Convolutional Neural Network [52]. A differential abstract soft attention component and additive difference in preference of the ROA pooling mechanism is used by Johnson et al[48]. This remodeling network chooses swimming supports in disseminate strategy and operating regions through. This sequence uses a sequence dataset for experiments to generate image caption field level.

A delineation of the whole optical locale is personalized and not sufficient to catalyze the whole perspective. Area-based elaboration is additional alternative and is border than world picture descriptions. Based on region elaboration can be considered as dense captioning. The congested caption has some challenges. An object may consist of an overlap area because the regions are dense. Furthermore, it is very strenuous to accept every targeted area for all optical ideas.

Yang et al. [53] planned additional densely translation technique. This technology will take care of these difficulties. Firstly, it present the an elation technique that simultaneously counts on the optical options of the area and also the anticipated tags for the that area. This enables the protype to seek the an acceptable situation of the delimiting box. Secondly, they implement a medium amalgamation to combine reference option along with optical choices from multiple fields to produce a chic linguistics description.

- *Caption for Total View:*

Architecture based on Encoder-Decoder, Integrative design, Attention-based architecture, Linguistics Architecture based on concepts, artificial caption Image captioning based on Novel object, and image captioning Methods based on Alternative Deep Learning Networks for or Multiple Caption total views.

- *Encoder-Decoder Architecture Vs. Compositional Architecture*

Few approaches simply make use if vanilla encoder and decoder directly or get subtitles. Nevertheless, alternative asks use multiple networks for this.

Image captioning based on encoder-decoder architecture: Image captioning methods based on Neural Network simply serve as way to eliminate straightforward methods. These strategies are almost like those of the encoder-decoder framework-based neural MT[73]. During this network, world picture options are mined from the secluded activities of CNN so they are fed into the An LSTM to obtain a sequel of words.

A standard approach for this classification has the subsequent common steps:

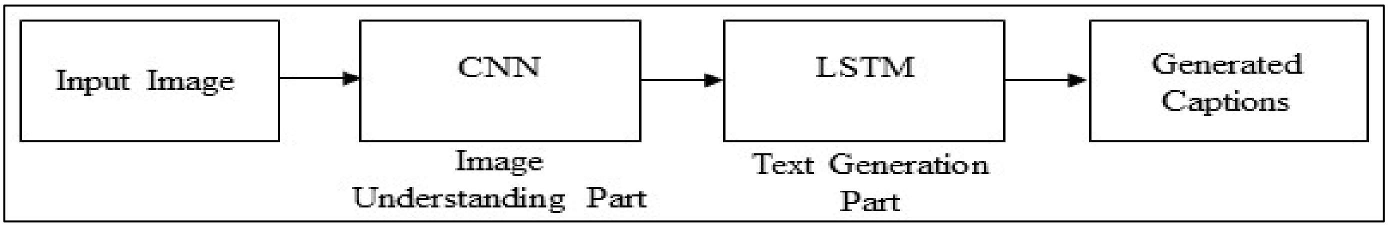


Fig 3: Simple Encoder- Decoder Architecture-Based Image Captioning

Figure shows a schematic protype of simple encoder-decoder architecture-based image captioning. Vinyals et al. planned a technique referred to as the Neural Image Caption Generator (NIC). For image representations and an LSTM to generate image captions the device make use of CNN. This particular CNN uses a completely unique technique for block standardization as well as production of the final hidden layer of CNN employed as the an input of the LSTM decoder. The LSTM is adept for tracking objects that have already destroyed the text that has been victimized. NIC is trained most likely the estimate has been supported.

In creating an image caption, the image data is appended to the primary state of the LSTM. Subsequent words are procreated that sustain the present time step and the preceding concealed condition. This method will be continued until it gets the tip indication of the statement. Since the picture data is feed completely at the commencement of the technique, it should encounter sequential issues that are missing. The role of initially generated words is additionally changing to weak and weak. Therefore, LSTM faces challenges in creating long-length sentences. As a result, Jia et al. [Extension [] the AN extension of the schematic LSTM is known as directed LSTM (gLSTM). This s gLSTM will produce longer sentences. Throughout the design, it connects.

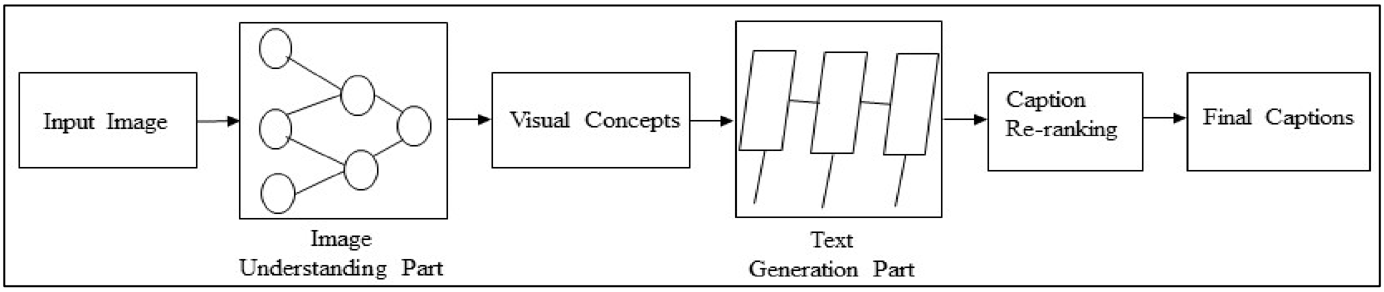


Fig 4: A Block Diagram of a Compositional Network-Based Captioning

Global linguistics data for every cell position of LSTM. Additionally, it considers completely diverse length standardization methods for regulating caption length. Linguistics data is excerpted in many conducts. Firstly, it utilize a sensory extraction function to obtain image captions so linguistics data is exerted from these tags. To employ a multimodal at embedding house most data-based linguistic s can be devised.

Mao et al. [68] planned a particular sort of text production technique for pictures. The technique will procreate an outline for AN precise object or area known as a referenced term [59, 60, and 64]. After falling victim to this expression, it will detect the thing or area that is delimited. Thus, the type of explanation or expression generated is ambiguous. Therefore to handle referring expressions, this technique uses a replacement d dataset referred In as the Refer It dataset [64] that supports well-liked MSCOCO datasets.

Most of the preceding CNN-RNN based image captioning techniques utilizes LSTM which is simpler rind comparatively low. In the simplex language production techniques, a successful word is required in a supported visual context and in every previous context. Simplex LSTM cannot procreate contextually well-fashioned tags. In addition, current object recognition and categorization techniques [65, 72] suggest that profound, hierarchical techniques are higher in erudition than shallow ones. Wang ct al. [Deep dup] planned a deep-duplex LSTM -based techniques for image captions. This technique is able to produce applicable rich image tags. The planned design contains one CNN and two isolated LSTM. It will use each precedent and upcoming reference data to detect optical verbal communication over a longer period of time.

- *Image Captioning is based on Compositional Architecture.*

Integrated architecture-based methods collected of numerous freelance purposeful structure blocks: Preliminary, a CNN is employed to extort linguistics ideas out of picture. After that a prototype based on language will be employed for obtaining a collection of candidate tags. During procreation of ultimate caption, these candidate tags are again employed on a deep multimodal similarity model.

➢ *Specific Technique for these Classes Maintains the Subsequent Steps:*

- CNN is employed by Image options.
- Visual choices derive Visual ideas (such as attributes).
- Several captions capture knowledge of step one and a pair of steps by a language model.

The procreated tags again employ a creep multimodal resemblance model to take prime superiority image captions. Generation-based image captioning was introduced by Fang et aI. [58]. It utilizes a optical detector, a model based on language, and a multimodal that coaches the prototype onto a picture tagging dataset. The picture caption will enclose nouns, verbs and adjectives. A terminology captures a thousand specific words from a coaching caption. The structure mechanisms with image sub-fields rather full image. Interactive neural networks (both AlexNet and VGG16NET are utilized to extract options for subfields of a picture. Subclasses options are mapped with vocabulary words that may possibly be contained within image captions are done. Multiple Instances Learning (MIL) is employed to coach the model for learning discriminative visual signatures of every word. A most entropy (ME) language prototype has been employed to generate image tags from these words. The procreated tags are categorized by the linear weighting of statement choices. To detect these weights Minimum error rate coaching (MERT) [71] is engaged. A common vector depiction is employed to calculate the resemblance between image and sentence. There is a mapping of image and sentence fragments by a deep multimodal similarity model (DMSM) with a normal vector illustration. This acquires a major advancement in selecting key excellence figure tags.

Several major methods have so far achieved acceptable improvement in image caption making. As of continuous domains training and testing samples are utilized by techniques. As a result, there is no conference that these prototypes will execute fine in open-domain images. In addition, they are fully sensible in identifying common visual content. There are some prominent organizations such as Celebrity and Landscape which are beyond their scope. Captions of those techniques are executed on automatic metrics such as cheese, METEOR [1] and potable. The examination metrics contains previously revealed rational results on these techniques. Conversely, an external distinction exists between the analysis of the matrix in terms of performance and the human judgment of analysis [66]. If this has been thought of as real-world entity data, performance may well he weak. Conversely, Tran et al.

[Distinct 4] introduced a diverse image captioning technique. This technique is also able to procreate image tags for open domain images scene will look at different sets of ideas and produce captions for celebrities and landmarks. It utilizes an peripheral mental entity freebase to identify individual entities like celebrities and landmarks. A sequence of human judgments is practical to evaluate the performance of the procreated tags. In experiments, it uses three datasets: MS Coco, Adobe-MIT FiveK [57], and pictures from Instagram. Pictures of the MS COCO dataset were composed from the continuous domain, but photographs from the alternative dataset were selected from the an open domain. The strategy achieves distinguished performance on a particularly difficult Instagram dataset.

A different integrative network-based image captioning technique was planned by Ma et al. [67]. This technique utilizes structural words <object, attribute, activity, and scene> to get meaningful explanation. It additionally utilizes a multi-task technique almost like several instance learning technique [58], and multi-layer improvement technique to get structural words. AN LSTM encoder- decoder-based MT technique is then accustomed interpret the structural terms to picture caption.

A parallel-fusion RNN-LSTM design for image caption production was designed by Wang et al. [78]. The design of the strategy segregates the concealed units of RNN and LSTM into a diversity of elements of similar dimension. Elements pair parallel with equivalent ratios to get hold of image tags.

- *LSTM Vs Others*

Image captioning intersects PC idea and linguistic communication process (NLP) analysis. NLP functions, in common, often develop as series education sequences. Numerous neural language prototypes such as the neural probabilistic language prototype [56], the log-bilinear model, the skip-gram model [70], and the perennial neural network (RNN) are planned for the sequence of learning sequence functions. RNN has been widely used a variety of series learning works. Nevertheless, ancient RNNs undergo extinction and gradient issues and cannot sufficiently switch semi-permanent chorological dependence.

LSTM [62] is a type of network RNN that consists of particular unit that are in accumulation to plain units. A memory cell is employed by LSTM which can retain data in your memory for a long time. In the current period, models based on LSTM are used in order to progression learning work. An additional network, the gated perennial unit (GRU), includes a structure related to the LSTM, although it does not use partitioned memory cells to control the stream of information (data) and uses fewer gates.

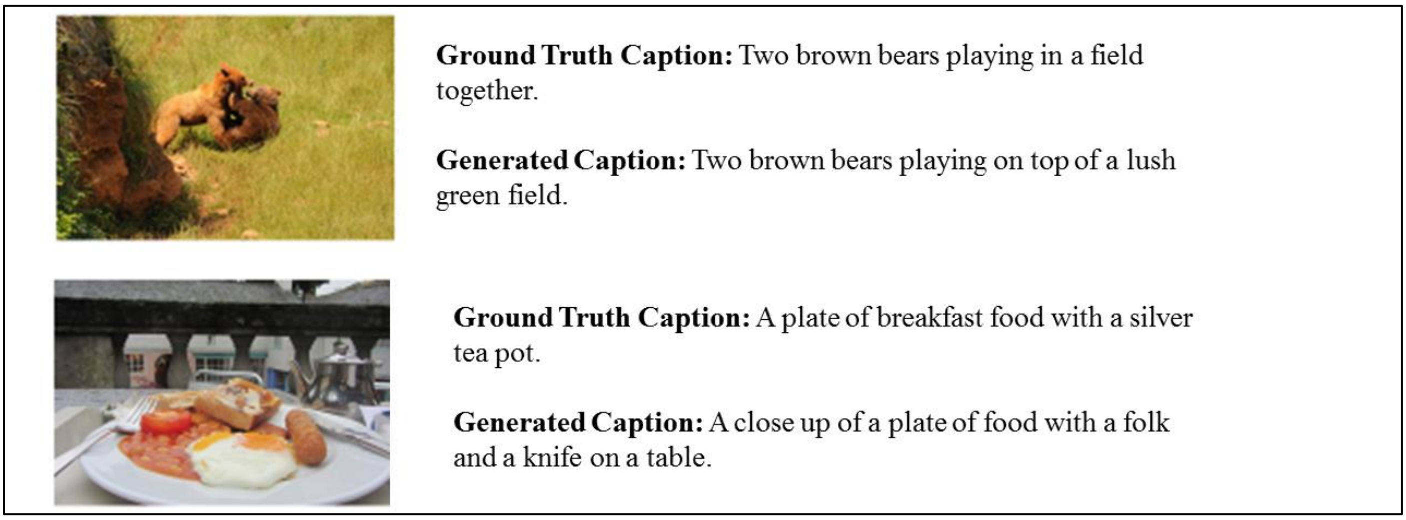
Nonetheless, the fundamental hierarchical data structure of a sentence is overlooked by LSTMs. Furthermore, they necessitate considerable storage due to a semi-permanent dependency through the memory cell. In CNNs will be more rapid in process than LSTM as they know the internal hierarchical data structure of sentences. As a result, in recent times, alternative functions are used in alternative sequences for succession functions, for example, conditional image formation [75] and MT.

More motivated by the achievement of CNNs in progression learning tasks, Gu et al. [41] planned a image caption technique based on CNN language model. The technique utilizes one verbal communication CNN for applied mathematical model based on language. Nonetheless, the strategy is not able to employ the dynamic chronological actions of the prototype as simply a language-CNN. It temporarily amalgamates a perennial network with language-CNN to properly prototype dependency. Aneja et al. [5] planned a rigorous design for the work of image tagging. They utilize feed-forward networks in which none of the perennials function. Tactic's design consists s of four components: (i) the input embedding layer (ii) the image embedding layer (iii) the concentric module, and (iv) the output embedding layer. In addition, it utilizes consideration mechanisms to take advantage of abstract image options. They estimate their design on hard MSCOCO datasets and show equivalent presentation to most LSTM-based techniques on customary matrix.

Wang et at. [102] planned an additional CNN + CNN that mostly feature picture caption technology. This is related to the strategy of Aneja et al. Apart from that to engage sight-CNN with language-CNN it utilizes a hierarchical concentration component. The use of a variety of hypermeters as well as the number of layers and the kernel width of the language-CNN are additionally investigated by the authors. They show that the effect of hyper parameters will advance strategy presentation in image tagging.

III. DATASETS AND EVALUATION METRICS

Several collections of data are used for training , examining, and analysis of image tagging techniques. The dataset takes a variety of perspectives such as the number of images, the number of captions per image, the arrangement of the caption, and the dimension of the image. 3 datasets: Flickr8k [87], Flickr30k [96], and MSCOCO datasets [93] are prevalently used. These collections of data along with others are given in division. Datasets and evaluation metrics Section 4.2 discussed the Metrics



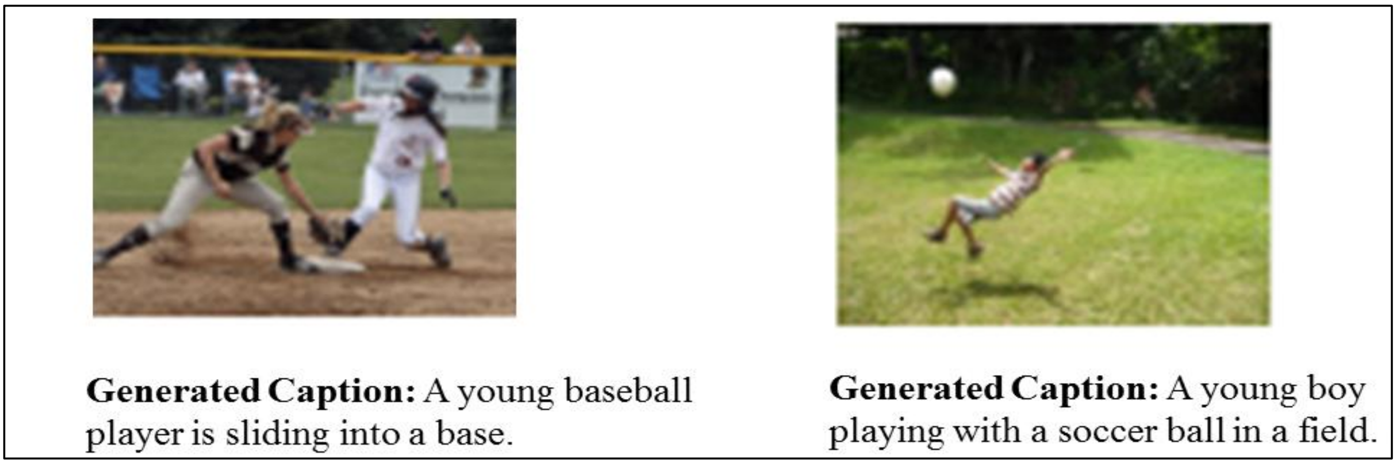
Ground Truth Caption: Two brown bears playing in a field together.

Generated Caption: Two brown bears playing on top of a lush green field.

Ground Truth Caption: A plate of breakfast food with a silver tea pot.

Generated Caption: A close up of a plate of food with a folk and a knife on a table.

Fig 5: Captions Generated by Wu et al. [103] on Some Sample Images from the MS COCO Dataset



Generated Caption: A young baseball player is sliding into a base.

Generated Caption: A young boy playing with a soccer ball in a field.

Fig 6: Captions generated by Chen et al. [84] on Some Sample Images from the Flickr30k Dataset

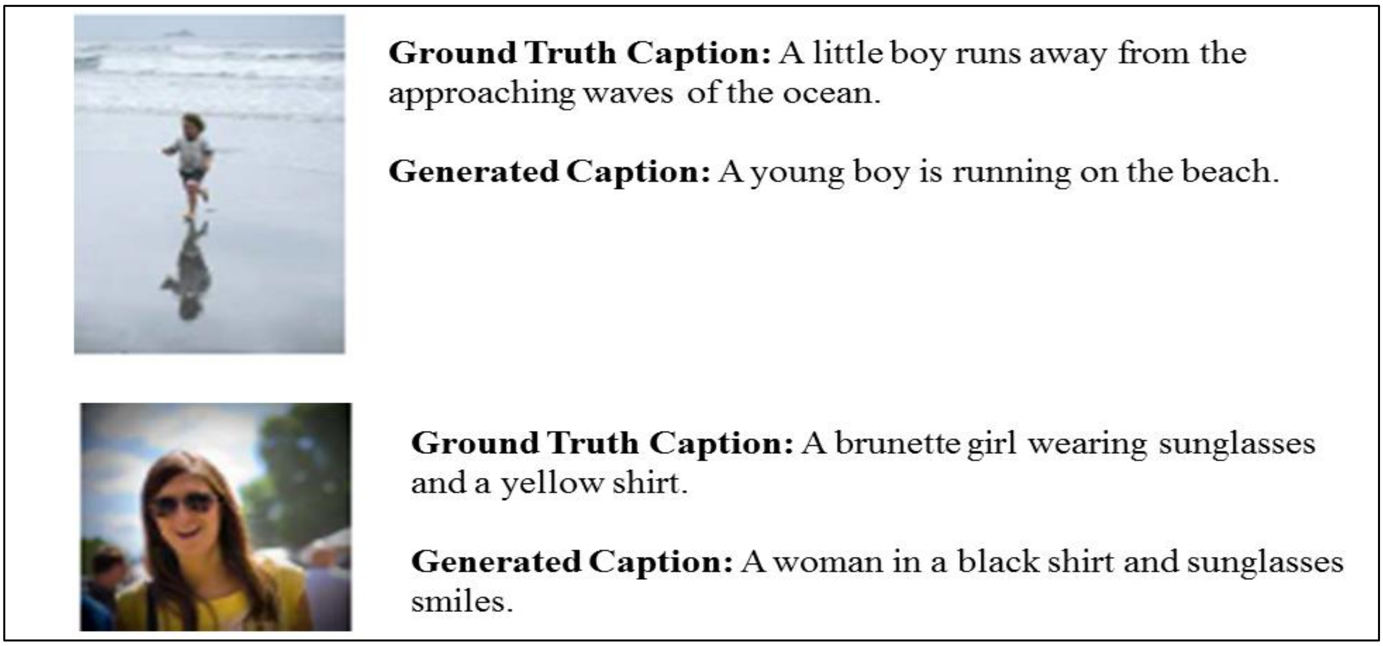
A. Datasets

➢ *MSCOCO Dataset*

The Microsoft COCO dataset [93] can be an exceptionally huge dataset for image recognition, captioning, and segmentation. The MS COCO dataset has various options such as object segmentation, context validation, numerous objects per category, quite three hundred, images, and a pair of million instances, eighty object classes, and five captions per image. Several images tagging techniques [101, 103, 104, and 105] use datasets in their practical. For example, Wu et at. [103] employ the MSCOCO dataset in their technique, as well as the captions of the 2 sample images shown in Figure 6.

➢ *Flickr30K Dataset*

For programmed image explanation and ground language consideration Flickr30k [96] dataset can be used. It contains 30,000 images gathered from Flickr with l58,000 tags provided by human annotators. It does not present any mounted partitioning of pictures for training, examining and verification. Number for training, testing, and verifications are chosen by the scholars themselves. The collection of data (dataset) has a detector for general objects, a color classification, and a partiality towards choosing larger objects. Techniques based on Image Caption such as [Image 490, 100] utilize this dataset for his or her experiments . For example, Flickr did its experiment on a 30k dataset. Caption procreated by the bird genus et al. [[4] the 2 sample datasets tire shown in Fig use 12.



Ground Truth Caption: A little boy runs away from the approaching waves of the ocean.

Generated Caption: A young boy is running on the beach.

Ground Truth Caption: A brunette girl wearing sunglasses and a yellow shirt.

Generated Caption: A woman in a black shirt and sunglasses smiles.

Fig 7: Captions Generated by Jia et al. [63] on Some Sample Pictures from the Flickr8k Dataset

➢ *Flickr8K Dataset*

Flickr 8k [87] can be a well-liked dataset and can contain 8000 photos gathered from Flickr. The coaching knowledge contains of 6k drawing s, checks and advancement knowledge, each consisting of 1k pictures. Each image within the dataset has five suggestion captions annotate by humans. A variety of techniques based on image captioning [83, 88] have made experiments a victim of the dataset. 2 sample results by Jia et at. Figure 7 shows [[Are] Pictures on this dataset.

- *Dataset on Visual Order*

The visual order dataset [91] is an additional dataset for image tagging. Image captioning does not need to identify only the items of a photograph, although it additionally argues their communications and characteristics. Not like the primary 3 dataset where the caption is given for the total view, a sequence in the scene sequence dataset has different captions for several regions. There are seven main parts of the dataset: field explanation, items, attributes, relations, field graphs, visual graphs and question answer pairs. The dataset contains quite 108,000 images. Every image has a meaning of 35 items, 2h features, and 21 pair wise associations between items.

- *Dataset on Instagram*

Two dataset victim images from Instagram which can be a photo-sharing social networking service were created by Tran et at. [98] and Park et al. [95]. The dataset created by Tran et al. which consists of 10k pictures is mainly from celebrities. However, Park et at. The social media network used its dataset for hash tag forecast and post-procreation tasks. This dataset contains a. 1m post on many good topics and a detailed hash tag list for over half a dozen users.

- *Dataset on IAPR TC-12*

The dataset of IAPR TC-12 [86] contains 20,000 pictures. Pictures are composed from diverse sources such as sports, pictures of individuals, animals, landscapes and lots of alternative places in the world. Captions in several languages have been given in photographs of this dataset. Pictures also contain many objects.

- *Dataset on Stock3M*

The dataset on Stock3M contains 217,654 photos uploaded by users. The dataset is 26 times larger than the MSCOCO dataset. The pictures in this datasets are diverse.

- *MIT-Adobe FiveK Dataset*

The MIT-Adobe FiveK [81] dataset contains 5000 pictures. The paintings consist of scene collection, subjects, and lighting situation and are primarily concerned with human, nature, and artificial items.

- *Flickr Sryle10k Dataset*

The FlickrStyle 10k dataset contains 10000 Flickr images with artificial tagging. The knowledge of coaching includes 7000 images. The 2000 and 1,000 illustrations are critically involved in verification and check knowledge. Every picture has romantic, comic and realistic captions.

B. Evaluation Metrics

- *BLEU*

BLEU (Bilingual Analysis Interpretation) [94] can be a metric that is accustomed to the standard of device procreated text. Scores are calculated for each individual text section comparing them with group of reference texts. Iri calculating the general feature of the produced text, the calculated score average. Nevertheless, syntax accuracy is not thought of here. The presentation of the Paneer metric is diverse, given the amount of orientation translation and the range of the produced text. Later, Papineni et al. introduced a changed accuracy metric. These matrixes utilize n-grams. Cheese is well liked as a result of being a pioneer in automated analysis of machine translated text and includes a cheap correlation with human judgments of quality [82, 85]. However, there are some limitations such as the cheese scores are sensible which the generated text is concise [82]. The standard of the generated text is not always considered sweet with the enhancement in cheese score for some cases.

- *ROUGE*

Recall-Oriented Understanding for Grouting Evaluation (ROUGE) [92] can be a set of metrics used to standardize text outlines. With a Collection of indication summaries produced by humans it compares word sequences, word pairs, and n-grams. Intended for different functions different types like ROUGE-1, 2, ROUGE-W, ROUGE-SU4 are used. For instance, ROUGE- I and ROUGE We are acceptable for distinct article analysis while ROUGE-2 and ROUGE-SU4 have sensible presentation in a brief summary. However, when it comes to multi-document text outlines, ROUGE has trouble in examining.

- *METEOR.*

METEOR (Express Ordering for Analysis of Translation with Metric) [5] Another metric accustomed is the use of machine-translated language. The customary word clauses are compared to reference texts. Furthermore, for the present, synonyms of a sentence stem and words are also thought of. METEOR will create a high correlation at the sentence or section level.

- *CIDEr*

Consensus-based Image Description Evaluation (CIDEr) [99] is an automatic agreement metric for examining image explanation. The majority of the presented datasets have exclusively five captions per image. Subsequent analysis metrics work with this tiny range of sentences and don't seem to be sufficient to live the agreement among generated captions and human judgment. On the other hand, potable achieves human agreement victimization Term Frequericy -Inverse Document Frequency (TF-IDF)

- *SPICE*

Semantic Propositional Image Caption Evaluation (SPICE) [3] could be a novel tag analysis metric supported Linguistic construct. It's supported a graph -based linguistics illustration referred to as scene-graph [89, 97]. This graph will extract the knowledge of various objects, attributes, and their relationships from the image descriptions.

IV. DISCUSSION

A. Comparison of Competition Models

A survey based on deep neural networks, we are presenting a great deal of competition models and automated value metrics, we are going to do some experiments to quantitatively match and analyze the model. Pictures and related captions of the square measurements are shown in fig. 5a and 5b. Table one has high performance in the harmonization attention model, with caption generation of the manipulation, and similar performance in the linguistic consideration prototype. In the linguistics attention prototype, the top-down and bottom-up methods square measure joint to extract rich feature data. Therefore, worldwide, native substitute square measurements for caption generation are well integrated. Additionally, to overview the associate image, linguistics attention models add rich visual idea as rich in granularity to give the outline an additional look and correct as Figure 5a. As on example, "a type of food" originates in the NIC model, while in the ATT- FCN model is represented by "sandwiches and French fries", similarly "eating a portion of paper” and the N cell model. As 'a cell phone “. In ATT-FCN model the correct description comprised to 'a toothbrush in his mouth" and "a combination of scissors ". However, the original choices extracted may be objective for unrelated visa features (or fewer visual attributes) that may constrain the model to elicit incorrect views. The visual feature "clock" shown in fig. 5A can guide the prototype to concentrate on the conditions data and to overlook the most foreground data to generate an explanation. In intense captioning, a quick tag for every identified item is created whether it is foreground or background. In the reconciliation focus model will advance the precision of explanation and positioning shown in fig. 5a and 5b. Sometimes, this produces incorrect captions. Samples of cohesion failure in the attention model class measurements shown in fig. 5 b. corresponding to instances of failure, though square measure focuses on proper captioning.

Table 1: Performance Testing of Models by the Online MS-COCO Test

Metrics Models					
	Template- Based MS captivator [15]	**Google NIC [53]**	**Hard attention [66]**	**Semantic attention ATT-FCN [68]**	**Adaptive attention [75]**
B-1					
c5	0.715	0.713	0.705	0.731	0.746
c40	0.907	0.895	0.881	0.9	0.918
B-2					
c5	0.543	0.542	0.528	0.565	0.582
c40	0.819	0.802	0.779	0.815	0.842
B-3					
c5	0.407	0.407	0.383	0.424	0.443
c40	0.71	0.694	0.658	0.709	0.74
B-4					
c5	0.308	0.309	0.277	0.316	0.335
c40	0.601	0.587	0.537	0.599	0.633
METEROR					
c5	0.248	0.254	0.241	0.25	0.264
c40	0.339	0.346	0.322	0.335	0.359
ROUGE-L					
c5	0.526	0.53	0.516	0.535	0.55
c40	0.68	0.682	0.654	0.682	0.706
CIDEr					
c5	0.931	0.943	0.865	0.943	1.037
c40	0.937	0.946	0.893	0.958	1.051

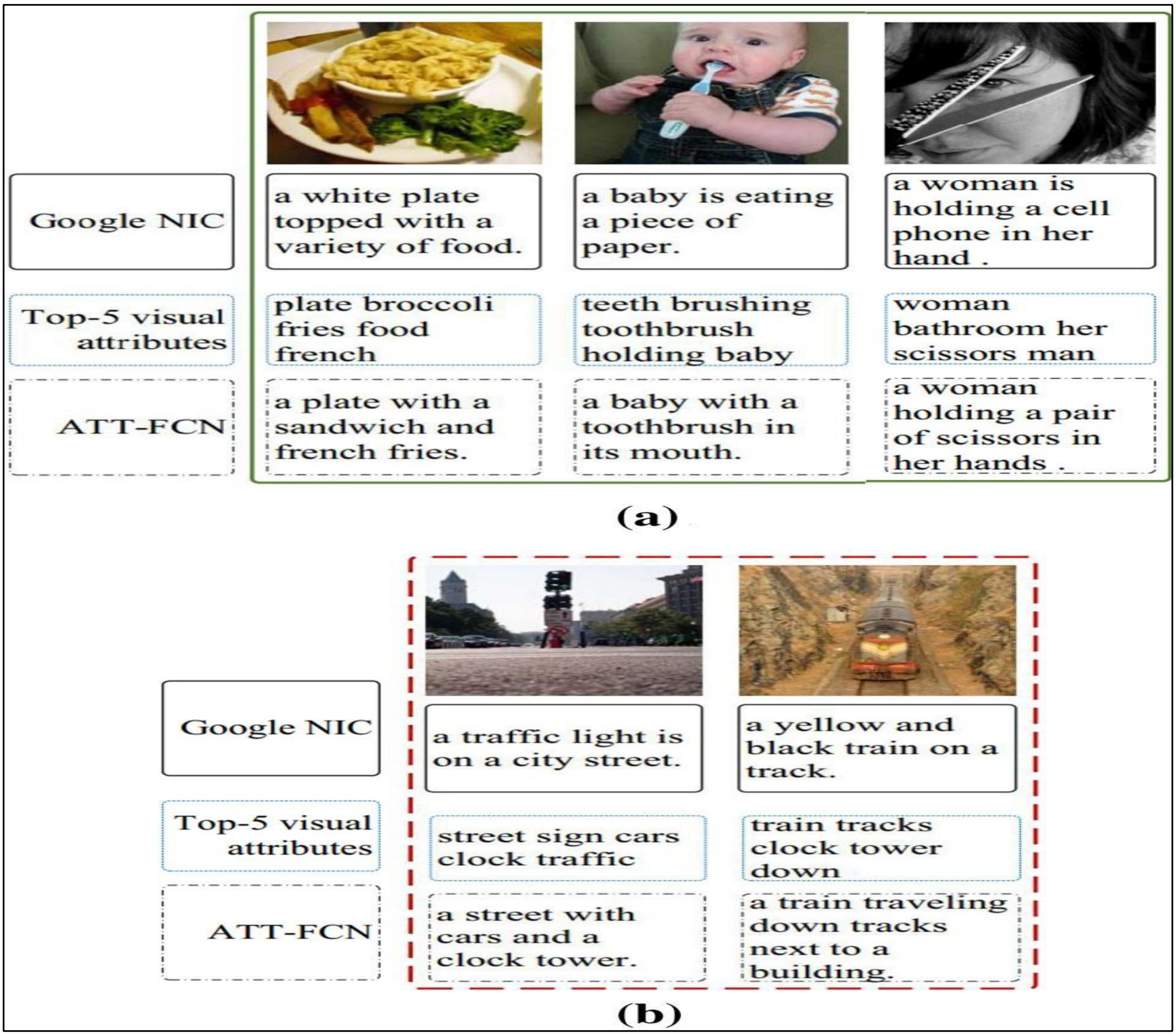

Google NIC
Top-5 visual attributes
ATT-FCN
a white plate topped with a variety of food.
plate broccoli fries food french
a plate with a sandwich and french fries.
a baby is eating a piece of paper.
teeth brushing toothbrush holding baby
a baby with a toothbrush in its mouth.
a woman is holding a cell phone in her hand .
woman bathroom her scissors man
a woman holding a pair of scissors in her hands .
(a)
Google NIC
Top-5 visual attributes
ATT-FCN
a traffic light is on a city street.
street sign cars clock traffic
a street with cars and a clock tower.
a yellow and black train on a track.
train tracks clock tower down
a train traveling down tracks next to a building.
(b)

Fig. 8: Comparisons of Two Competing Models. a Positive Contrast and b Reverse Contrast Areas of the Image, the Model Generates Inaccurate Descriptions which are Caused by the Error of Texture Recognition

V. CONCLUSIONS

In this paper, we have reviewed in -depth learning-based image captioning methods. We have given an assortment of image captioning techniques, showing general pictures of the main teams and highlighting their pros and cons. We refer to completely different analysis matrices and datasets with their strengths and weaknesses. A short outline of the experimental results is additionally given. We go into the short printed potential analysis directions during this space. Although intensive learning-based image captioning techniques has acquired a stimulating advancement in current time, a robust image captioning technique that is determined to produce captions of prime quality for almost all images is however to be acquired. With the presence of original intensive network architectures, automated image captioning may continue an energetic analysis space over a period of time.

REFERENCES

[1]. Kuznetsova, P., Ordonez, V., Berg, A.C., Berg, T.L., Choi, Y.: Collective generation of natural image descriptions. In: Meeting of the Association for Computational Linguistics: Long Papers, Korea, Jeju Island, pp. 359–368 (2012)

[2]. Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng, A.Y.: Grounded compositional semantics for finding and describing images with sentences. Trans. Assoc. Comput. Linguist. 2, 207–218 (2014)

[3]. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep Boltzmannmachines. J.Mach. Learn. Res. 15, 2949–2980 (2014)

[4]. Norouzi, M.,Mikolov,T., Bengio, S., Singer,Y., Shlens, J., Frome, A., Corrado, G.S., Dean, J.: Zero-shot learning by convex combination of semantic embeddings. In: International Conference on Learning Representations ICLR2014, Banff, Canada (2014)

[5]. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: data, models and evaluation metrics. J. Artif. Intell. Res. 47, 853–899 (2013)

[6]. Ordonez,V.,Kulkarni, G.,Berg, T.L.: Im2Text: describing images using 1 million captioned photographs. Adv. Neural Inf. Process. Syst. 25, 1143–1151 (2012)

[7]. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252 (2014)

[8]. Li, S., Kulkarni, G., Berg, L.B., Berg, C.A., Choi, Y.: Composing simple image descriptions using web-scale N-grams. In: 15th Conference on Computational Natural Language Learning, Portland, USA, 2011, pp. 220–228 (2011)

[9]. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., Zitnick, C.L., Zweig, G.: From captions to visual concepts and back. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, MA, USA, pp. 1473–1482 (2015)

[10]. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: International Conference on Neural Information Processing Systems, MIT, Montreal, pp. 3104–3112 (2014)

[11]. Johnson, R., Zhang, T.: Effective Use of Word Order for Text Categorization with Convolutional Neural Networks, pp. 103–112. Eprint Arxiv arXiv:1412.1058 (2014)

[12]. Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473 (2014)

[13]. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: 2015 Computer Vision and Pattern Recognition, IEEE, Boston, pp. 3156–3164 (2015)

[14]. Channarukul, S., Mcroy, S.W., Ali, S.S.: DOGHED: a templatebased generator for multimodal dialog systems targeting heterogeneous devices. In: Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (2003)

[15]. Oriol, V., Alexander, T., Samy, B., Dumitru, E.: Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge. IEEE Trans Pattern Anal. 39(2017), 652–663 (2015)

[16]. Ioffe, S., Szegedy, C., Bach, F., Blei, D.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: 32nd International Conference on Machine Learning, ICML 2015, International Machine Learning Society (IMLS), Lille, France, pp. 448–456 (2015)

[17]. Mao, J., Xu,W., Yang, Y.,Wang, J., Yuille, A.L.: Explain Images with Multimodal Recurrent Neural Networks. arXiv preprint arXiv:1410.1090 (2014)

[18]. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Darrell, T., Saenko, K.: Long-term recurrent convolutional networks for visual recognition and description In: Computer Vision and Pattern Recognition, IEEE, Boston, MA, USA, p. 677 (2015)

[19]. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, arXiv preprint arXiv:1411.2539 (2014)

[20]. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, Attend and Tell: Neural Image CaptionGenerationwithVisual Attention, Computer Science, pp. 2048–2057 (2015)

[21]. You, Q., Jin,H.,Wang, Z., Fang, C.,Luo, J.: Image captioning with semantic attention. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Las Vegas, NV, USA, pp. 4651–4659 (2016)

[22]. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning, arXiv preprint arXiv:1612.01887 (2016)

[23]. Shuang Bai and Shan An. 2018. A Survey on Automatic Image Caption Generation. Neurocomputing.

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

[25]. Cristian Bodnar. 2018. Text to Image Synthesis Using Generative Adversarial Networks. arXiv preprint arXiv:1805.00676.

[26]. Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory. ACM, 144–152.

[27]. Xinlei Chen and C Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2422–2431.

[28]. Bo Dai, Dahua Lin, Raquel Urtasun, and Sanja Fidler. 2017. Towards Diverse and Natural Image Descriptions via a Conditional GAN. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2989–2998.

[29]. William Fedus, Ian Goodfellow, and Andrew M Dai. 2018. Maskgan: Better text generation via filling in the _. arXiv preprint arXiv:1801.07736.

[30]. Chuang Gan, Tianbao Yang, and Boqing Gong. 2016. Learning attributes equals multi-source domain generalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 87–97.

[31]. Ross Girshick. 2015. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision. 1440–1448.

[32]. Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 580–587.

[33]. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Advances in neural information processing systems. 2672–2680.

[34]. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778.

[35]. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (CVPR). 5967–5976.

[36]. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems. 1097–1105.

[37]. Akshi Kumar and Shivali Goel. 2017. A survey of evolution of image captioning techniques. International Journal of Hybrid Intelligent Systems Preprint, 1–19.

[38]. Yann LeCun, LÃľon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 11 (1998), 2278–2324.

[39]. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In International Conference on Learning Representations (ICLR).

[40]. Timo Ojala, Matti PietikÃďinen, and Topi MÃďenpÃďÃď. 2000. Gray scale and rotation invariant texture classification with local binary patterns. In European Conference on Computer Vision. Springer, 404–420.

[41]. Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In International Conference on learning Representations (ICLR).

[42]. Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In Proceedings of Machine Learning Research, Vol. 48. 1060–1069.

[43]. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems. 91–99.

[44]. Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training. In IEEE International Conference on Computer Vision (ICCV). 4155–4164.

[45]. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1–9.

[46]. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3156–3164.

[47]. Heng Wang, Zengchang Qin, and Tao Wan. 2018. Text Generation Based on Generative Adversarial Nets with Latent Variables. In Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 92–103.

[48]. Ross Girshick. 2015. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision. 1440–1448.

[49]. Sepp Hochreiter and JÃijrgen Schmidhuber. 1997. Long short-term memory. Neural computation 9, 8 (1997), 1735–1780.

[50]. Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4565–4574.

[51]. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision 123, 1 (2017), 32–73.

[52]. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems. 91–99.

[53]. Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. 2016. Dense Captioning with Joint Inference and Visual Context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 1978–1987.

[54]. Abhaya Agarwal and Alon Lavie. 2008. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. In Proceedings of the ThirdWorkshop on Statistical Machine Translation. Association for Computational Linguistics, 115–118.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1602

[55]. Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. 2018. Convolutional image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5561–5570.

[56]. Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. Journal of machine learning research 3, Feb, 1137–1155.

[57]. Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. 2011. Learning photographic global tonal adjustment with a database of input/output image pairs. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 97–104.

[58]. Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr DollÃąr, Jianfeng Gao, Xiaodong He, Margaret Mitchell, and John C Platt. 2015. From captions to visual concepts and back. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1473–1482.

[59]. Nicholas FitzGerald, Yoav Artzi, and Luke Zettlemoyer. 2013. Learning distributions over logical forms for referring expression generation. In Proceedings of the 2013 conference on empirical methods in natural language processing. 1914–1925.

[60]. Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In Proceedings of the 2010 conference on empirical methods in natural language processing. Association for Computational Linguistics, 410–419.

[61]. Jiuxiang Gu, GangWang, Jianfei Cai, and Tsuhan Chen. 2017. An empirical study of language cnn for image captioning. In Proceedings of the International Conference on Computer Vision (ICCV). 1231–1240.

[62]. Sepp Hochreiter and JÃijrgen Schmidhuber. 1997. Long short-term memory. Neural computation 9, 8 (1997), 1735–1780.

[63]. Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In Proceedings of the IEEE International Conference on Computer Vision. 2407–2415.

[64]. Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes.. In EMNLP. 787–798.

[65]. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems. 1097–1105.

[66]. Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Baby talk: Understanding and generating image descriptions. In Proceedings of the 24th CVPR. Citeseer.

[67]. Shubo Ma and Yahong Han. 2016. Describing images by feeding LSTM with structural words. In Multimedia and Expo (ICME), 2016 IEEE International Conference on. IEEE, 1–6.

[68]. Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 11–20.

[69]. Oded Maron and TomÃąs Lozano-PÃľrez. 1998. A framework for multiple-instance learning. In Advances in neural information processing systems. 570–576.

[70]. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[71]. Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 160–167.

[72]. Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations (ICLR).

[73]. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Advances in neural information processing systems. 3104–3112.

[74]. Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. 2016. Rich image captioning in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 49–56.

[75]. Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. 2016. Conditional image generation with pixelcnn decoders. In Advances in Neural Information Processing Systems. 4790–4798.

[76]. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3156–3164.

[77]. Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. 2016. Image captioning with deep bidirectional LSTMs. In Proceedings of the 2016 ACM on Multimedia Conference. ACM, 988–997.

[78]. Minsi Wang, Li Song, Xiaokang Yang, and Chuanfei Luo. 2016. A parallel-fusion RNN-LSTM architecture for image caption generation. In Image Processing (ICIP), 2016 IEEE International Conference on. IEEE, 4448–4452.

[79]. Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In European Conference on Computer Vision. Springer, 382–398.

[80]. Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, Vol. 29. 65–72.

[81]. Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. 2011. Learning photographic global tonal adjustment with a database of input/output image pairs. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 97–104.

[82]. Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the Role of Bleu in Machine Translation Research.. In EACL, Vol. 6. 249–256.

[83]. Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, and Tat-Seng Chua. 2017. SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 6298–6306.

[84]. Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang,Wan-Ting Hsu, Jianlong Fu, and Min Sun. 2017. Show, Adapt and Tell: Adversarial Training of Cross-domain Image Captioner. In The IEEE International Conference on Computer Vision (ICCV), Vol. 2.

[85]. Etienne Denoual and Yves Lepage. 2005. BLEU in characters: towards automatic MT evaluation in languages without word delimiters. In Companion Volume to the Proceedings of the Second International Joint Conference on Natural Language Processing. 81–86.

[86]. Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In International workshop ontoImage, Vol. 5. 10.

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

[88]. Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In Proceedings of the IEEE International Conference on Computer Vision. 2407–2415.

[89]. Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3668–3678.

[90]. Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3128–3137.

[91]. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision 123, 1 (2017), 32–73.

[92]. Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop, Vol. 8. Barcelona, Spain.

[93]. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr DollÃạr, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In European conference on computer vision. Springer, 740–755.

[94]. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 311–318.

[95]. Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to You: Personalized Image Captioning with Context Sequence Memory Networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 6432–6440.

[96]. Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision. 2641–2649.

[97]. Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In Proceedings of the fourth workshop on vision and language, Vol. 2.

[98]. Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. 2016. Rich image captioning in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 49–56.

[99]. Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 4566–4575.

[100]. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3156–3164.

[101]. Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. 2016. Image captioning with deep bidirectional LSTMs. In Proceedings of the 2016 ACM on Multimedia Conference. ACM, 988–997.

[102]. Qingzhong Wang and Antoni B Chan. 2018. CNN+ CNN: Convolutional Decoders for Image Captioning. arXiv preprint arXiv:1805.09019.

[103]. Qi Wu, Chunhua Shen, Anton van den Hengel, Lingqiao Liu, and Anthony Dick. 2015. Image captioning with an intermediate attributes layer. arXiv preprint arXiv:1506.01144 (2015).

[104]. Zhilin Yang Ye Yuan Yuexin Wu and Ruslan Salakhutdinov William W Cohen. 2016. Encode, Review, and Decode: Reviewer Module for Caption Generation. In 30th Conference on Neural Image Processing System(NIPS).

[105]. Quanzeng You, Hailin Jin, ZhaowenWang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4651–4659.7.